GENDER DIFFERENCES ON COLLEGE ADMISSION TEST ITEMS: EXPLORING THE ROLE OF MATHEMATICAL BACKGROUND AND TEST ANXIETY USING MULTIPLE METHODS OF DIFFERENTIAL ITEM FUNCTIONING DETECTION

By

THOMAS E. LANGENFELD

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL OF
THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1995

ACKNOWLEDGEMENTS

I would like to express my sincerest appreciation to the individuals who have assisted me in completing this study. I am extremely indebted to Dr. Linda Crocker, chairperson of my doctoral committee, for helping me in the conceptualization, development, and writing of this dissertation. Her assistance and encouragement were extremely important in enabling me to achieve my doctorate. I also want to thank the other members of my committee, Dr. James Algina, Dr. Jin-win Hsu, Dr. Marc Mahlios, and Dr. Rodman Webb, for patiently reading the manuscript, offering constructive comments, providing editorial assistance, and giving continuous support. I further wish to thank Dr. David Miller, Dr. John Hall, and Scott Behrens for their assistance related to different aspects of this study.

I want to express my deepest gratitude to my family for providing the emotional support that was so vital during my graduate experience. I want to thank my wife, Ann--in many ways this degree is as much hers as mine, for understanding and encouraging me during my graduate

studies. Space limitations do not allow me to express the many personal sacrifices made by my wife so that I could complete this study. I also want to thank my daughter, Kathryn Louise, who was born in the early stages of this study and has come to provide a special type of support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract of Dissertation Presented to the Graduate School of the University of Florida in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

GENDER DIFFERENCES ON COLLEGE ADMISSION TEST ITEMS: EXPLORING THE ROLE OF MATHEMATICAL BACKGROUND AND TEST ANXIETY USING MULTIPLE METHODS OF DIFFERENTIAL ITEM FUNCTIONING DETECTION

By

Thomas E. Langenfeld

August, 1995

Chairperson: Linda Crocker
Major Department: Foundations of Education

The purpose of this study was to discover whether defining examinee subpopulations by relevant educational or psychological variables, rather than gender, would yield item statistics that were more consistent across five methods for detection of differential item functioning (DIF). A subsidiary purpose of this study was to assess how the consistency of DIF estimates were affected when structural validation findings were incorporated into the analysis. The study was conducted in the context of college admission quantitative examinations and gender issues. Participants consisted of 1263 university students. For purposes of this study, their responses to a 30-item quantitative aptitude test were analyzed by categorizing examinees by their gender, mathematics backgrounds, and levels of test anxiety.

The hypothesis that defining subpopulations by mathematics background or test anxiety would yield higher consistency of DIF estimation than defining subpopulations by gender was not substantiated. Results indicated that using mathematics background to define subpopulations and explain gender DIF had potential usefulness; however, in this study, the use of test anxiety to define subpopulations and explain DIF was ineffectual.

The findings confirmed the importance of structural validation for DIF analyses. Results from using the entire test revealed that nonuniform DIF methods had low inter-method consistency and variance related to methods. When structural validation findings were used to define a valid subset of items, highly consistent DIF indices resulted across all methods and minimal variance related to methods. Results further suggested the need to use nonuniform DIF methods and the importance of jointly interpreting both DIF indices and significant tests. Implications and recommendations for research and practice are included.

CHAPTER 1
INTRODUCTION

Statement of the Problem

Differential item functioning (DIF), a statistical indication of item bias, occurs when equally proficient individuals, from different subpopulations, have different probabilities of answering an item correctly (R. L. Linn, Levine, Hastings, & Wardrop, 1981; Scheuneman, 1979; Shepard, Camilli, & Williams, 1984). Historically, researchers studying DIF have addressed two principal concerns. The first concern of researchers has been the development and evaluation of statistical methods for detecting "biased" items. The second concern has been to identify plausible explanations of item bias. In this study, both methodological and substantive educational issues concerning item bias and DIF were addressed.

During the past four decades, a plethora of detection methods has been developed (for a comprehensive review of advances in item bias detection methods over the past ten years see Millsap and Everson, 1993). DIF methods have two major distinctions: (a) whether the conditioning variable is formed from an observed conditional score or an unobserved conditional estimate of latent ability and (b) whether they can detect nonuniform as well as uniform DIF.

Researchers applying methods using an observed conditional score commonly sum the number of correct responses on the test or subsection of the test to estimate the ability of each examinee. Researchers using unobserved conditional estimates most frequently apply a unidimensional item response theory (IRT) model for estimating the latent ability of each examinee.

Uniform DIF occurs when there is no interaction between ability level and group membership. That is, the probability of answering an item correctly is greater for one group than the other group uniformly over all ability levels. Nonuniform DIF, detectable only by some methods, occurs when there is interaction between ability level and group membership. That is, the difference in the probabilities of a correct response for the two groups is not the same at all ability levels. In IRT terms, nonuniform DIF is indicated by "nonparallel" item characteristic curves.

Among the various DIF procedures, the Mantel-Haenszel (MH) statistic, as applied by Holland and Thayer (1988), has emerged as the most widely used procedure (more because of the Educational Testing Service's usage than as a result of theoretical consensus), and it is frequently the method to which others are compared (Hambleton & Rogers, 1989; Raju, 1990; Shealy & Stout, 1993a; Swaminathan & Rogers, 1990). The appeal of the MH procedure is its simple conceptualization, relative ease of use, chi-square test of significance, and desirable statistical properties (Dorans & Holland, 1993; Millsap & Everson, 1993). Researchers applying MH employ an observed score as the conditioning variable and recognize that MH is sensitive to only uniform DIF.

Other methods compared with the MH procedure in this study included logistic regression (Swaminathan & Rogers, 1990), IRT-Signed Area (IRT-SA), IRT-Unsigned Area (IRT-UA) (Raju, 1988, 1990), and the Simultaneous Item Bias Test (SIBTEST) (Shealy & Stout, 1993a, 1993b). Logistic regression was designed to condition on observed scores analyzing item responses. With logistic regression, the user can detect both uniform and nonuniform DIF. IRT-SA and IRT-UA were devised to condition on latent ability estimates and assess the area between an item characteristic curve (ICC) estimated for one subgroup against an ICC estimated for a second subgroup. IRT-SA was developed to detect only uniform DIF, whereas IRT-UA was developed to detect both uniform and nonuniform DIF. SIBTEST was designed to conceptualize DIF as a multidimensional phenomenon where nuisance determinants adversely influence item responses (Shealy & Stout, 1993a, 1993b). Researchers using SIBTEST apply factor analysis to define a valid subtest and a regression correction procedure to estimate the criterion variable. SIBTEST was developed to detect only uniform DIF.

In assessing different DIF indices with data from a curriculum-based, eighth grade mathematics test, Skaggs and Lissitz (1992) found that the consistency between methods was low, and no reasonable explanation for items manifesting DIF could be hypothesized. They posited that categorizing subpopulations by demographic characteristics such as gender or ethnicity for DIF studies was "not very helpful in conceptualizing cognitive issues and indicated nothing of the reasons for the differences" (p. 239). A number of researchers have suggested the need to explore DIF using subpopulations categorized by psychologically and educationally significant variables that correlate with gender or ethnicity and potentially influence item performance (R. L. Linn, 1993; Schmitt & Dorans, 1990;

Skaggs & Lissitz, 1992; K. K. Tatsuoka, R. L. Linn, M. M. Tatsuoka, & Yamamoto, 1988).

Thus, a major concern of the study was the consistency of results from different DIF estimation procedures when subpopulations are conceptualized by psychological or educational variables. Three methods of conceptualizing subpopulations were combined with five fundamentally different state-of-the-art procedures for assessing DIF.

The Measurement Context of the Study

The substantive issue was the investigation of gender differences on a sample test containing items similar to those found on advanced college admission quantitative examinations. Generally, men tend to outperform women on the Scholastic Aptitude Test-Math (SAT-M), the American College Testing Assessment Mathematics Usage Test (ACT-M), and the Graduate Record Examination-Quantitative (GRE-Q). However, from a predictive validity perspective, these differences are problematic. For example, men tend to score approximately 0.4 standard deviation units higher on the SAT-M (National Center for Education Statistics, 1993), although women tend to perform at nearly the same level as men in equivalent college mathematics courses (Bridgeman & Wendler, 1991; Wainer & Steinberg, 1992) and tend to outperform men in general college courses (Young, 1991, 1994).

A possible explanation for quantitative test score differences between men and women is background experience. Men tend to enroll in more years of mathematics (National Center for Education Statistics, 1993). A second explanation that could potentially explain the differential validity of such tests is test anxiety. Test anxiety relates to examinees' fears of negative evaluation and defensiveness (Hembree, 1988). Women generally report higher levels of test anxiety than men (Everson, Millsap, & Rodriquez, 1991; Hembree, 1988; Wigfield & Eccles, 1989). Thus, for high-stakes tests of mathematical aptitude, mathematics background and test anxiety could influence item responses differentially for each gender.

The Research Problem

In this study, I explored the feasibility of conceptualizing subpopulations by relevant educational or psychological variables in contrast to the use of traditional demographic variables. The vehicle to achieve this purpose was a released form of the GRE-Q. In this study, subpopulations were conceptualized by (a) gender groups, one of the traditional demographic groups, (b) examinees with substantial and little mathematics background, and (c) examinees high and low in test anxiety. DIF was assessed using five different measures. The DIF measures were MH, logistic regression, IRT-SA, IRT-UA, and SIBTEST. The DIF methods were classified into two groups--methods measuring uniform DIF and alternate methods. The uniform methods were MH, IRT-SA, and SIBTEST. Alternate methods included logistic regression and IRT-UA, along with MH. Logistic regression and IRT-UA were designed to measure both uniform and nonuniform DIF. Mantel-Haenszel was placed into both analysis groups because of its widespread use by test practitioners. Regarding the study's methodological issues, the results of five methods of estimating DIF will be contrasted within each of the three modes of defining subpopulation groups.

The observation of interest was the DIF indices estimated for each item under a particular combination of subpopulation definition and DIF method. Replications were the 30 items on a released form of the GRE-Q test. For the research questions that follow, trait effects refer to the three subpopulation conceptualizations and method effects refer to the five methods of estimating DIF indices.

The first four research questions address the consistency of DIF indices between methods when subpopulations are conceptualized using different traits. The uniform methods of MH, IRT-SA, and SIBTEST were combined with the traits gender, mathematics background, and test anxiety to yield a multitrait-multimethod (MTMM) matrix of correlation coefficients. (See Table 1 for an illustration of a MTMM matrix with uniform measures.) Similarly, the alternate DIF estimation methods of MH, IRT-UA, and logistic regression were combined with the traits gender, mathematics background, and test anxiety to yield a second multitrait-multimethod (MTMM) matrix of correlation coefficients. (See Table 2 for an illustration of a MTMM matrix with alternate measures.)

Each of the following research questions was addressed twice; each question was answered for the uniform methods and the alternate methods, respectively:

1. Among the three sets of convergent coefficients, often termed the monotrait-heteromethod coefficients, (e.g., the correlation between the DIF indices obtained from the MH and IRT-SA methods when subpopulations are defined by the trait gender), will the coefficients base upon the subpopulations of mathematics background or test anxiety be significantly larger than the corresponding

Table 1

Proposed Multitrait-Multimethod Correlation Matrix: Uniform DIF Indices

	MH-D			IRT-SA			SIBTEST-b		
	A	B	C	A	B	C	A	B	C
I.MH-D									
A.Gender	( )								
B.MathBkd	H-M	( )							
C.TA	H-M	H-M	( )						
II.IRT-SA									
A.Gender	M-H*	H-H	H-H	( )					
B.MathBkd	H-H	M-H*	H-H	H-M	( )				
C.TA	H-H	H-H	M-H*	H-M	H-M	( )			
III.SIBTEST-b									
A.Gender	M-H*	H-H	H-H	M-H*	H-H	H-H	( )		
B.MathBkd	H-H	M-H*	H-H	H-H	M-H*	H-H	H-M	( )	
C.TA	H-H	H-H	M-H*	H-H	H-H	M-H*	H-M	H-M	( )

Note. ( ) = reliability coefficients. M-H* = monotrait-heteromethod or the convergent validity coefficients. H-M = heterotrait-monomethod coefficients. H-H = heterotrait-heteromethod coefficient.

Table 2

Proposed Multitrait-Multimethod Correlation Matrix: Alternate DIF Indices

	MH-D			IRT-UA			Log Reg		
	A	B	C	A	B	C	A	B	C
I.MH-D									
A.Gender	( )								
B.MathBkd	H-M	( )							
C.TA	H-M	H-M	( )						
II.IRT-UA									
A.Gender	M-H*	H-H	H-H	( )					
B.MathBkd	H-H	M-H*	H-H	H-M	( )				
C.TA	H-H	H-H	M-H*	H-M	H-M	( )			
III.Log Reg									
A.Gender	M-H*	H-H	H-H	M-H*	H-H	H-H	( )		
B.MathBkd	H-H	M-H*	H-H	H-H	M-H*	H-H	H-M	( )	
C.TA	H-H	H-H	M-H*	H-H	H-H	M-H*	H-M	H-M	( )

Note. ( ) = reliability coefficients. M-H* = monotrait-heteromethod or the convergent validity coefficients. H-M = heterotrait-monomethod coefficients. H-H = heterotrait-heteromethod coefficient.

coefficients when subpopulations are defined by gender?

2. Will the monotrait-heteromethod coefficients be higher than the coefficients for different traits measured by the same method (i.e., heterotrait-monomethod coefficients)?

3. Will the convergent correlation coefficients be higher than the discriminant coefficients measuring different traits by different methods(i.e., heterotrait-heteromethod coefficients)?

4. Will the patterns of correlations among the three traits be similar over the three methods of DIF estimation?

The final research question addressed the consistency of DIF procedures in identifying aberrant items when subpopulations are conceptualized in different ways. The question was applied twice; it was answered for the uniform methods and alternate methods. It was as follows:

5. For each DIF detection method respectively, using standard decision rules, what is the percent agreement about aberrant items when subgroups are based on gender and when subgroups are based on (a) mathematics background and (b) test anxiety?

Following the analysis of uniform and alternate DIF methods, I conducted a structural analysis of the 30-item quantitative test. Shealy and Stout (1993a, 1993b) stressed that practitioners must carefully identify a valid subset of items prior to conducting DIF analyses. They argued that DIF occurs as a consequence of multidimensionality. The potential for DIF occurs when one or more nuisance dimensions interact with the valid dimension of a test (Ackerman, 1992; Camilli, 1992). Messick (1988) stressed the structural component of construct validation. The structural component concerned the extent to which items are combined into scores that reflect the structure of the underlying latent construct. Loevinger (1957) termed the purity of the internal relationships as <u>structural fidelity</u>, and it is appraised by analyzing the interitem structure of a test. I employed factor analytic procedures to define a structurally valid subset of unidimensional items and to identify problematic and multidimensional items. I hoped to define items measuring both the intended dimension and nuisance.

After the identification of a structurally valid subset of items, I assimilated the findings and repeated the DIF analysis described above. Again, I assessed DIF using the five methods with subpopulations defined by gender, mathematics background, and test anxiety. Using DIF indices as the unit of analysis, two MTMM matrices of correlation coefficients were generated--one matrix for uniform methods and one matrix for alternate methods. I applied the five research questions to the MTMM matrices and inferential statistics using the structurally valid subset of items. I contrasted the findings of the analysis for the entire test with the findings of the analysis for the subset of test items.

Theoretical Rationale

The process of ensuring that high-stakes tests contain no items that function differentially for specific subpopulations is a fundamental concern of construct validation. Items that contain nuisance determinants correlated with an examinee subpopulation membership threaten the construct interpretations derived from test scores for that subpopulation. Psychometric researchers continue to examine the merits of numerous DIF detection procedures and explore theoretical explanations of DIF. However, to date, they have failed to reach consensus on methodological issues or to develop meaningful insight concerning its causes. In part, this failure is the consequence of inconsistent DIF identification with actual test data (R. L. Linn, 1993; Shepard et al., 1984; Skaggs & Lissitz, 1992). These concerns were investigated from both a practical and theoretical perspective that has been suggested (R. L. Linn, 1993; Schmitt & Dorans, 1990; Skaggs & Lissitz, 1992; K. K. Tatsuoka et al., 1988) but rarely tested.

Two significant premises underlie the study. The first premise is that there is nothing inherent in being a female examinee or a member of a specific ethnic group that predisposes an individual to find a particular item troublesome. Educational and psychological phenomena function in unique ways to disadvantage an individual on a specific item. Traditional DIF occurs when phenomena correlate with the demographic group of interest. Consequently, gender or ethnicity can be interpreted as a surrogate for educational or psychological variables that potentially explain DIF's causes.

Skaggs and Lissitz (1992) posited that educational and psychological variables that influence item performance and correlate with ethnic or gender groups would be useful for conceptualizing subpopulations. Millsap and Everson (1993) commented that modeling variables such as educational background and test anxiety might assist in understanding DIF's causes. The educational and psychological variables in the study that were hypothesized as potentially explaining gender DIF on quantitative test items were mathematics background and test anxiety.

Mathematics background was selected because it influences quantitative reasoning and problem solving. Further, high school and college men tend to enroll in more mathematics courses and study mathematics at more abstract levels than women (National Center for Educational Statistics, 1993). Researchers assessing overall SAT-M performance have found that gender differences decrease substantially when differences in high school mathematics background are taken into account, although background does not entirely explain score differences (Ethington & Wolfle, 1984; Fennema & Sherman, 1977; Pallas & Alexander, 1983). Quantitative aptitude test scores can be contaminated by familiarity with item context, the application of novel solutions, and the use of partial knowledge to solve complex problems (Kimball, 1989). These types of skills frequently are developed through background experiences.

Test anxiety was selected because of its well-documented debilitating influence on examinees' performance. (Hembree, 1988; Hill & Wigfield, 1984;

Liebert & Morris, 1967; Tryon, 1980). For individuals possessing high levels of test anxiety, test scores frequently are depressed and construct interpretations become problematic (Everson et al., 1991; Hembree, 1988; Sarason, 1980). Consequently, test anxiety exemplifies a psychological variable that potentially contaminates the construct interpretations of scores. For examinees with high levels of test anxiety, tests of mathematical ability tend to induce extreme levels of anxiety. (Richardson & Woolfolk, 1980). Female students tend to report higher levels of test anxiety than male students at all grade levels including college (Everson et al., 1991; Hembree, 1988; Wigfield & Eccles, 1989). Over the past 20 years, several self-reported measures of test anxiety have been developed that demonstrate high reliability and well-defined theoretical properties (Benson, Moulin-Julian, Schwarzer, Seipp, & El Zahhar, 1991; Sarason, 1984; Spielberger, Gonzalez, Taylor, Algaze, & Anton, 1978). Researchers have used the self-reported instruments to measure test anxiety and assess the efficacy of treatment programs (Sarason, 1980; Spielberger et al., 1978; Wine, 1980). For studying gender DIF on college admissions quantitative items, it was hypothesized that test anxiety possessed strong explanatory potential because of its threat to valid score interpretation, negative influence on tests of mathematics ability, and gender effects.

A second premise underlying the study is a fundamental tenet of educational measurement. Item responses are products of complex interactions between examinees and a set of items. In part, because of this complex interaction, examinees of approximately equivalent abilities who belong to different subpopulations occasionally have different likelihoods of answering a question correctly. This fascinating finding currently is understood only crudely. Before it can be better understood, the effect of DIF detection methods and different means of conceptualizing subpopulations on item responses must be examined.

Limitations of the Study

A salient limitation of the study was the nature of the performance task. Participants in the study were administered a sample GRE-Q and were told they would have 30 minutes to complete the test. They were told to perform to the best of their ability and they would be able to learn their results following testing. Although every effort was made to simulate the conditions of a high-stakes examination, if participants felt that the performance had little meaning for them, then their performance might not accurately reflect their performance on a high-stakes college admissions test. Further, if the participants believed that the examination had low-stakes, the levels of test anxiety felt by examinees while answering the sample GRE-Q would not be equivalent to levels of test anxiety experienced by examinees while answering a college admissions test.

Finally, examinees in the study were predominantly undergraduate students taking classes in the colleges of education and business at a large, southern state university. For this reason, although the design, methodology, and analysis were conceived and executed to maximize the generalizability of findings, a degree of caution is recommended in generalizing to other populations or settings.

CHAPTER 2
REVIEW OF LITERATURE

The four central aspects of this study were Differential Item Functioning (DIF) methodology, gender differences in mathematical college-level aptitude testing, gender differences in mathematics background, and test anxiety. These four topics constitute the major themes for the organization of the literature review presented in this chapter.

DIF Methodology

A Conceptual Framework of DIF

Tests for placement in education and selection in employment require scores be fair and representative for all individuals. Since the mid-1960s, measurement specialists have been concerned explicitly with the fairness of their instruments and the possibility that some tests may be biased (Cole & Moss, 1989). Bias studies initially were designed to investigate the assertions that disparities between various subpopulations on cognitive ability test scores were a product of cultural bias inherent in the measure (Angoff, 1993). Test critics charged that bias occurred whenever mean scores for two subpopulations were not equivalent. This position accepted a priori that subpopulations had equivalent score distributions on the construct measured and dismissed the possibility that actual differences may exist. Measurement specialists, however, have resolved that mean score differences do not necessarily reflect bias but indicate test impact (Dorans & Holland, 1993).

Concerns about measurement bias are inherent to validity theory (Cole & Moss, 1989). A test score inference is considered sufficiently valid when various types of evidence justify its usage and eliminate other counterinterpretations (Messick, 1989; Moss, 1992). Bias has been characterized as "a source of invalidity that keeps some examinees with the trait or knowledge being measured from demonstrating that ability" (Shepard, Camilli, & Williams, 1985, p.79). If score-based inferences are not equally valid for all relevant subgroups, decisions derived from score inferences will not be fair for all individuals. Therefore, measurement bias occurs when score interpretations are differentially valid for any subgroup of test takers (Cole & Moss, 1989).

To investigate the potential for measurement bias, researchers have examined test items as a source and explanation. The supposition is that biased items require knowledge and skills that examinees from a specified subgroup are less familiar with and possess fewer opportunities to learn (Angoff, 1993). The goals of item bias research are to identify and remove items detected as biased (Angoff, 1993) and to provide test developers with guidelines making future construction of biased items less likely (Scheuneman, 1987; Schmitt, Holland, & Dorans, 1993).

Measurement specialists have defined item bias as occurring when individuals, from different subpopulations, who are equally proficient on the construct measured have different probabilities of successfully answering the item (Angoff, 1993; R. L. Linn, Levine, Hastings, & Wardrop, 1981; Scheuneman, 1979; Shepard et al., 1985). Researchers apply statistical methods to equate individuals on the construct, utilizing either observed scores or latent ability scores, and estimate for examinees of each group the probability of a correct response. These methods provide statistical evidence of bias. When a statistically biased item is identified, it might be interpreted as unfairly disadvantageous to a minority group for cultural and social reasons. On the other hand, the item might be interpreted as unrelated to cultural and social factors but related to an important educational outcome that is not equally known and understood by all groups. In this latter case, deleting the item for strictly statistical reasons may reduce validity.

Consequently, social and statistical definitions of item bias have created considerable confusion within the debate over test fairness (Cole & Moss, 1989; Angoff, 1993).

Researchers discovered that statistical analyses of item bias raised expectations and created confusion for an already obscure and volatile topic. The term _differential item functioning_ (DIF) gradually has replaced item bias as the preferred term in research because of its more neutral and technical connotations (Angoff, 1993; Dorans & Kulick, 1986). Holland and Wainer (1993) distinguished between item bias and DIF stating, item bias refers to "an informed judgment about an item that takes into account the purpose of the test, the relevant experiences of certain subgroups of examinees taking it, and statistical information about the item" (p. xiv). DIF is a "relative term" (p. xiv) and is a statistical indication of a differential response pattern. Shealy and Stout (1993a) proposed that the difference between item bias and DIF is "the degree the user or researcher has embraced a construct validity argument" (p. 197).

Shealy and Stout (1993a, 1993b) conceptualized DIF as a violation of the unidimensional nature of test items. They classified the intended dimension as the _target ability_ and unintended dimensions as _nuisance determinants_. DIF occurred because of nuisance determinants existing in differing degrees among subgroups. Crocker and Algina (1986) postulated that DIF occurred if (a) for subgroups, items are affected by different sources of variance; and (b) among test takers who are at the same point on the construct, the distributions of irrelevant sources of variation are different for subgroups. Therefore, DIF can be conceptualized as a consequence of multidimensionality with differing sources of variation influencing subgroups' item responses.

A Formal Definition of DIF

All DIF detection methods rely on assessment of response patterns of subgroups to test items. The subgroups, conceptualized in most studies on the basis of demographic characteristics (i.e., blacks and whites, women and men), form a categorical variable. When two groups are contrasted, the group of interest (e.g., blacks or women) is designated the focal group, and the group serving as the group for comparison (e.g., whites or men) is designated the reference group. Examinees are matched on a criterion variable, assumed to be a valid representation of the purported construct, and DIF methods assess differential group response patterns for individuals of equal ability.

Denote the item score as Y, frequently scored as a dichotomous variable 0 or 1; denote X as the conditioning criterion; and denote g as the categorical variable of group membership. Lack of measurement bias or DIF for an item is defined as

$$P_g(Y=1 \mid X) = P_{g'}(Y=1 \mid X)$$

for all values of X for the reference and focal groups. In this definition, $P_g(Y=1|X)$ is the conditional probability function for Y at all levels of X (Millsap & Everson, 1993).

Although all DIF procedures operate from this definition, they differ on the basis of statistical models and possess various advantages. DIF procedures can be characterized as models using observed conditional invariance or models utilizing unobserved conditional invariance (Millsap & Everson, 1993). When observed conditional invariance is used, the criterion variable is the sum of the total number of correct responses on the test or a subset of the test. When unobserved conditional invariance is used, a unidimensional item response theory (IRT) model estimates a θ parameter for each examinee that functions as the criterion variable.

Other differences in DIF detection procedures are the capacity to detect nonuniform DIF, to test statistical significance, and to conceptualize DIF as a consequence of multidimensionality. Uniform DIF occurs when there is no interaction between group membership and the conditioning criterion regarding the probability of answering an item correctly. In other words, DIF functions in a uniform fashion across the ability spectrum. Nonuniform DIF refers to an interaction between group membership and the conditioning criterion. In this case, an item might differentially favor a subgroup of examinees at one end of the ability spectrum and disfavor the subgroup at the other end of the spectrum. All DIF procedures are used to estimate an index describing the magnitude of the differential response pattern for the groups on the item. Some procedures also provide statistical tests to detect if the DIF index differs significantly from zero. Finally, although DIF is perceived as a consequence of multidimensionality, every procedure except Shealy and Stout's Simultaneous Bias Test (SIBTEST) functions within an unidimensional framework.

Many DIF detection methods have been developed during the past three decades. In this review, they are categorized as based upon observed conditional invariance or unobserved latent conditional invariance. Related issues, research problems, and potential usage are evaluated. Following the review of DIF detection methods, research efforts to explain the underlying causes of DIF are presented.

DIF Methods Based Upon Observed Scores

Angoff and Ford (1973) offered the first widely used DIF detection method called the delta-plot. The delta-plot procedure was problematic due to its tendency, under conditions of differing ability score distributions, to identify the most discriminating items as aberrant (Angoff, 1993). Scheuneman (1979) proposed a chi-square procedure for assessing DIF. This procedure was irrelevantly affected by sample size and was not based upon a chi-square sampling distributions, in effect, not a chi-square procedure at all (Baker, 1981). The full chi-square procedure (Bishop, Fienberg, & Holland, 1975) was a valid technique for testing DIF but required large sample sizes at each ability level to sustain statistical power. Holland and Thayer (1988) built upon these chi-square techniques when they applied the Mantel and Haenszel (1959) statistic, originally developed for medical research, to the detection of DIF.

Mantel-Haenszel procedure. The Mantel-Haenszel (MH) statistic has become the most widely used method of DIF detection (Millsap & Everson, 1993). The MH procedure assesses the item data in a J-by-2-by-2 contingency table. At each score level j, individual item data are presented for the two groups and the two levels of item response, right or wrong (see Table 3).

The null hypothesis for the MH procedure can be expressed as the odds of answering an item correctly at a given ability level are the same for both groups across all j ability levels. The alternative hypothesis is that the two groups do not have equal probability of answering the item correctly at some level of j.

Table 3

Item Data for the 2 Groups and 2 Item Scores for the jth Ability Group

		Score on Studied Item		
		1	0	Total
Group	R	A_{Rj}	B_{Rj}	n_{Rj}
	F	C_{Fj}	D_{Fj}	n_{Fj}
	Total	m_{1j}	m_{0j}	T_j

The MH statistic uses a constant odds ratio (α_{MH}) as an index of DIF. The estimate of the constant odds ratio is

$$\alpha_{MH} = \frac{[\sum_{j=1}^{J} A_{Rj} D_{Fj} /T_j]}{[\sum_{j=1}^{J} C_{Fj} B_{Rj} /T_j]} .$$

The constant odds ratio ranges in value from zero to infinity. The estimated value of α_{mh} is 1 under the null condition. It is interpreted as the average factor by which the odds that a reference group examinee will answer the item correctly exceeds that of a focal group examinee. Consequently, an estimated constant odds ratio greater than 1 indicates the case where the item is functioning differentially against the focal group.

The estimated value of α_{mh} frequently is transformed to the more easily interpreted Δ metric via

$$MH\ D\text{-}DIF = -2.35 \ln[\alpha_{MH}] .$$

Positive values of MH D-DIF favor the focal group, whereas negative values favor the reference group.

The chi-square test of significance for MH is

$$MH - \chi^2 = \frac{[|\sum_{j=1}^{J} A_{Rj} - \sum_{j=1}^{J} E(A_{Rj})| - .5]^2}{\sum_{j=1}^{J} Var(A_{Rj})},$$

where

$$E(A_{Rj}) = n_{Rj} m_{1j} / T_j$$

and

$$Var(A_{Rj}) = \frac{[n_{Rj} m_{1j} n_{Fj} m_{0j}]}{[T_j^2 (T_j - 1)]} .$$

The MH chi-square is distributed approximately as a chi-square with one degree of freedom. Holland and Thayer (1988) asserted that this test is "the uniformly most powerful unbiased test of H_o versus H_a" (p. 134).

The advantages of MH are its computational simplicity (Holland & Thayer, 1988), statistical test of significance, and lack of sensitivity to subgroup differences in the distribution of ability (Donoghue, Holland, & Thayer, 1993; Shealy & Stout, 1993a; Swaminathan & Rogers, 1990). The most frequently cited disadvantage is its lack of power to detect nonuniform DIF (Swaminathan & Rogers, 1990). It is further limited by its unidimensional conception and assumption that total test score provides a meaningful measure of the construct purported to be estimated.

The standardization procedure. The standardization procedure (Dorans & Kulick, 1986) is based upon the nonparametric regression of test scores on item scores for two groups. Let $E_R(Y|X)$ define the expected item test nonparametric regression for the reference group, and let $E_F(Y|X)$ define the expected item test nonparametric regression for the focal group, where Y is the item score and X is the test score. The DIF analysis at the individual score level is

$$D_j = E_{Fj} - E_{Rj} .$$

The statistic, D_j, is the fundamental measure of differences in item performance between the focal and reference group members who are matched at equivalent observed score levels. These differences are unexpected differences and cannot be explained by differences in the attribute tested.

The standardization procedure derived its name from the standardization group that functions to supply a set of weights, one at each ability level, that will be used to weight each individual D_j. The standardized p-difference (*STD P-DIF*) is

$$STD\text{-}P = \frac{\sum_{j=1}^{J} W_j \, (E_{Fj} - E_{Rj})}{\sum_{j=1}^{J} W_J},$$

or

$$STD\text{-}P = \frac{\sum_{j=1}^{J} W_j \, D_J}{\sum_{j=1}^{J} W_j}.$$

The essence of standardization is the weighting function. The specific weight implemented for standardization depends upon the nature of the study (Dorans & Kulick, 1986). Plausible options of weighting include the number of examinees in the total group at each level of j, the number of examinees in the focal group at each level of j, or the number of examinees in the reference group at each level of j. In practice, the number of examinees in the focal group is used; thus, STD-P is defined as the difference between the observed performance and the expected performance of the focal group on an item (Dorans & Kulick, 1986).

The standardization procedure contains a significance test. The standard error using focal group weighting is

$$SE(STD-P)=\sqrt{\frac{P_F(1-P_F)}{N_F}+VAR\ (P_F^*)},$$

where P_F is the proportion of focal group members correctly answering the item, and where P_F* is thought of as the performance of the focal group members predicted from the reference group's item test regression curve and

$$VAR\ (P_F^*)=\sum_{j=1}^{J}\frac{N_{Fj}^2 P_{Rj}\ (1\ -\ P_{Rj})}{N_{Rj}\ N_F^2}.$$

The standardization procedure is a flexible method of investigating DIF (Dorans & Holland, 1993), and it has been applied to assessing differential functioning distractors (Dorans, Schmitt, & Bleistein, 1992) and the differential effect of speededness (Schmitt & Dorans, 1990). DIF findings from the standardization procedure will be in close agreement with the MH procedure (Millsap & Everson, 1993) with the choice of weighting creating slight variations (Dorans & Holland, 1993). Because the two procedures are nearly identical, the advantages and disadvantages for the standardization method are much the same as for MH. The most commonly cited deficiency of both methods is their inability to detect nonuniform DIF. Donoghue et al. (1993) determined that both methods require approximately 19 or more items in the conditioning score, the studied item should be included in determining the conditioning score, and extreme ranges in item difficulty can adversely influence DIF estimation. R. L. Linn (1993) observed that DIF estimates using these procedures appear to be confounded with item discrimination.

Logistic regression models. Swaminathan and Rogers (1990) applied logistic regression to DIF analysis. Logistic regression models, unlike least squares regression, permit categorical variables as dependent variables. Thus, it permits the analysis of dichotomously scored item data. It has additional flexibility by including the analysis of interaction between group and ability, as well as allowing for the inclusion of other categorical and continuous independent variables in the model.

A fundamental concept of analysis with linear models is the assessment of the consistency between a model and a set of data (Darlington, 1990). Consistency between the model and the data set is measured by the likelihood or probability that the model correctly represents the observed data. When the dependent variable is measured dichotomously, scored 0 or 1, a model asserts that each examinee will have a probability between 0 and 1 of answering an item correctly. By the multiplicative law of independent probabilities, an overall probability for a group of examinees answering in a specific pattern can be estimated. For example, if the probability of four individuals each answering an item correctly is 0.9, and three of the subjects answer correctly, the overall probability of this pattern occurring is 0.9 X 0.9 X 0.9 X (1-0.9) or 0.0729. Therefore, for item i, the likelihood function of a set of examinee responses each with ability level θ is determined by

$$L(Data|\theta)=\prod_{n=1}^{N} P(u_i)^{u_i}[1-P(u_i)]^{1-u_i},$$

where u_i has a value of 1 for a correct response and a value of 0 for an incorrect response.

The logistic regression model for predicting the probability of a correct answer is

$$P(u=1|\theta) = \frac{\exp(\beta_0+\beta_1\theta)}{[1+\exp(\beta_0+\beta_1\theta)]},$$

where u is the response to the item given the ability level θ, β_o is the intercept parameter, and β_1 is the slope parameter. If categorical group variable g is added to the model for the analysis of DIF, the model becomes

$$P(u = 1|\theta) = \frac{\exp(\beta_0 + \beta_1\theta + \beta_2 g + \beta_3\theta g)}{[1 + \exp(\beta_0 + \beta_1\theta + \beta_2 g + \beta_3\theta g)]},$$

where β_2 is the estimate of uniform difference between groups, and β_3 is the estimated interaction between group and ability. If only β_0 and β_1 deviate from zero, the item is interpreted as containing no DIF. If β_2 does not equal zero, and β_3 equals zero, uniform DIF is indicated. If β_3 does not equal zero, nonuniform DIF is inferred.

Estimation of the parameters β_0, β_1, β_2, and β_3 is carried out for each item using a maximum likelihood procedure. The two null hypotheses can be tested jointly by

$$\chi^2 = \beta' C' (C\Sigma\, C)^{-1} C\beta,$$

where

$$C = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The test has a chi-square distribution with 2 degrees of freedom. When the test is significant, the hypothesis of no DIF is rejected (Swaminathan & Rogers, 1990).

The logistic regression procedure offers a powerful approach for testing the presence of both uniform and nonuniform DIF. In sample sizes of 250 and 500 examinees per group and with 40, 60, and 80 test items serving as the criterion, Swaminathan and Rogers (1990) concluded under conditions of uniform DIF that the logistic regression procedure had power similar to the MH procedure and controlled Type I errors almost as well. The logistic regression procedure also had effective power in identifying nonuniform DIF, whereas the MH procedure was virtually powerless to do so. In demonstrating the ineffectiveness of the MH procedure to detect nonuniform DIF, Swaminathan and Rogers (1990) simulated data keeping item difficulties equal and varying the discrimination parameter. In effect, they simulated nonuniform symmetrical DIF. Their simulation created a set of conditions where theoretically the MH procedure has no power. Researchers must ask whether such symmetrical interactions occur with actual test data. Millsap and Everson (1993) commented that Swaminathan and Rogers (1990) utilized large numbers of items, and they conjectured that in cases with a small number of homogeneous items forming the criterion variable, false positive rates would increase unacceptably above nominal levels.

Under conditions of uniform and nonuniform DIF, logistic regression facilitates the plotting of data and examining the differential response patterns to determine at which ability levels DIF is observed. The logistic procedure, although developed from a unidimensional perspective, provides a flexible model that can incorporate a diversity of independent categorical and continuous variables. Millsap and Everson (1993) observed that the procedure "allows for the inclusion of curvilinear terms and other factors--such as examinee characteristics like test anxiety or instructional opportunity--that may be relevant factors for exploring the possible causes of DIF" (p. 306).

DIF Methods Based Upon Latent Ability Estimation

DIF detection methods conditioning on latent ability are developed through various IRT models. IRT approaches describe the relationship between individual item responses and the construct measured by a test or subtest. When applied to DIF analyses, IRT permits the use of estimates of true ability as the criterion variable as opposed to the more subjective measure of observed scores. Despite its theoretical appeal, IRT approaches possess the inherent disadvantages of requiring large sample sizes, being computationally complex and costly, and including the stringent assumption of unidimensionality (Oshima, 1989). The most widely-used IRT models are the Rasch model or one-parameter model, the two-parameter logistic model (2PL), and the three-parameter logistic model (3PL). Holland and Thayer (1988) demonstrated that when the Rasch model's assumptions are met and none of the items constituting the ability score, except possibly the studied item, contain DIF, MH provides a DIF index proportional to the index estimated by the Rasch model. Therefore, methods based upon the Rasch model will not be reviewed, and the more complex 2PL and 3PL models will be reviewed regarding their potential.

The central components of IRT models are the unobserved latent trait estimate, termed θ, and a trace line for each item response, often termed the item characteristic curve (ICC). The ICC will take a specified monotonically increasing function. In the 2PL model, the probability of a correct response to Item i as a function of θ is

$$P(u_i = 1 | \theta) = \frac{\exp\,[Da_i\,(\theta - b_i)]}{1 - \exp\,[Da_i\,(\theta - b_i)]}\ ,$$

where the item parameters a_i and b_i are item discrimination and difficulty, respectively, and D is a constant set at 1.7 in order to convert the logistic scale into an approximate probit scale (Hambleton & Swaminathan, 1985). In the 3PL model, the probability of a correct response is

$$P(u_i = 1 | \theta) = c_i + (1 - c_i)\frac{\exp\,[Da_i\,(\theta - b_i)]}{1 - \exp\,[Da_i\,(\theta - b_i)]}\ ,$$

and includes the pseudo-chance or guessing parameter, c_i.

The general procedure for estimating DIF using a 3PL IRT model includes (a) combining both groups and estimating item parameters utilizing either a maximum likelihood or Bayesian procedure, (b) fixing the c_i parameter for all items, (c) after dividing the examinees into reference and focal group members, estimating the a_i and b_i parameters, (d) equating the parameters from the focal group scale to the reference group scale or vice versa, (e) calculating the DIF index and significance test, and (f) utilizing a purification procedure (Lord, 1980; Park & Lautenschlager, 1990) to further examine and enhance the analysis. Purification procedures, which extract potential DIF items and reestimate ability levels without the potential DIF items included, will not be elaborated upon. DIF indices and statistical tests based upon latent ability proceed by either analyzing the difference between the item parameters (a_i, b_i) or analyzing the area between the groups' ICCs.

Lord's chi-square and IRT-LR. Lord's (1980) chi-square and IRT-Likelihood Ratio (IRT-LR) simultaneously tests the dual hypothesis of $a_{Ri} = a_{Fi}$ and $b_{Ri} = b_{Fi}$. Because the pseudo-chance parameter and its standard errors are not accurately estimated for the separate groups (Kim, Cohen, & Kim, 1994), it is usually not tested with either procedure.

For a single parameter, Lord's chi-square contrasts the difference between the estimated b's and their standard errors. Since sample sizes used in IRT estimation are so large to effectively assume an infinite number of degrees of freedom, the test becomes

$$z = \frac{(b_R - b_F)}{\sqrt{var(b_R) + var(b_F)}} \, .$$

Alternately, z^2 will be distributed as a chi-square statistic with one degree of freedom (Thissen, Steinberg, & Wainer, 1988). The simultaneous test of the discrimination and difficulty parameters is based upon Mahalanobis distance (D^2) between the parameter vectors for the groups. The test statistic becomes

$$D^2 = V' \Sigma^{-1} V \ ,$$

in which **V** is the vector of differences between the parameter estimations (a_R - a_F and b_R - b_F) and Σ is the estimated covariance matrix. The test is distributed as a chi-square with two degrees of freedom.

The same hypothesis tested by Lord's chi-square can be tested with IRT-LR (Thissen, Steinberg, & Wainer 1993). The null hypothesis with IRT-LR is tested through three steps. The IRT model is fitted simultaneously for both groups to the data. A set of valid "anchor" items, containing no DIF, are utilized as the conditioning variable. In the first step, no constraints are placed on the data concerning the equality of the b's or a's. The model fit is assessed by maximum likelihood statistics and

$$G_1^2 = -2(loglikelihood).$$

The IRT model is refitted under the constraint that the b and a parameters are equal for both groups and

$$G_2^2 = -2(loglikelihood).$$

The likelihood ratio test of significance is the difference between the fit of the two models and is

$$G^2(2) = G_2^2 - G_1^2 .$$

The likelihood ratio test assesses significant improvement in model fit as a consequence of allowing the two parameters to fluctuate. If the likelihood ratio is significant, either the b parameter or the a parameter is different for the two groups, and DIF is detected. In this example, of simultaneously testing for differences in both parameters, the test statistic is distributed as a chi-square with two degrees of freedom. In the situation of testing for significance only in item difficulty, the statistic would be distributed as a chi-square with one degree of freedom.

Both Lord's chi-square and IRT-LR assume multivariate normality. The two procedures differ in the estimation of the covariance matrix. Lord's chi-square is based upon second-derivative approximations of the standard errors of estimated item parameters as a part of the maximum likelihood estimation. The IRT-LR procedure does not require estimated error variances and covariances. It results from computing the likelihood at the overall mode under the equality constraints placed upon the data and then estimating the probability under the null hypothesis (Thissen et al., 1988).

Lord's chi-square and IRT-LR are capable of detecting nonuniform DIF and possess good statistical power (Cohen & Kim, 1993). Because of the requisite large sample sizes, they tend to be expensive and yield false positive rates above the nominal levels (Kim et al., 1994). R. L. Linn et al. (1981), with simulated data, and Shepard, Camilli, and Williams (1984), with actual data, demonstrated that significant differences detected by Lord's chi-square occurred even when plotted ICCs were nearly identical. An additional problem when employing IRT-LR is the need for a set of truly unbiased anchor items (Millsap & Everson, 1993).

Procedures estimating area between ICCs. Eight different DIF procedures have been developed to estimate the area between the reference group's ICC and the focal group's ICC. Area measures differ by whether they employ (a) signed or unsigned areas, (b) bounded or unbounded ability intervals, (c) continuous integration or discrete approximation, and (d) weighting (Millsap & Everson, 1993).

The first area procedures utilized bounded intervals with discrete approximations. Rudner (1977) suggested an unsigned index

$$R = \sum_{\theta = -3}^{3} [\, |P_R(\theta_j) - P_F(\theta_j)|\, \Delta\,] \,,$$

with discrete intervals from $\theta_j = -3$ to $\theta_j = 3$. Rudner (1977) used small interval distances (e.g.; $\Delta = .005$) and summed across 600 intervals. The estimated R is converted to a signed index by removing the absolute value operator.

Shepard et al. (1984) extended the signed and unsigned area procedures by introducing four techniques that included sums of squared values, weights based upon number of examinees in each interval along the θ scale, and weighting initial differences by the inverse of the estimated standard error of the difference. They determined that distinctively different interpretations occurred when signed area indices were estimated as compared to unsigned indices. They further found that various weighting procedures influence interpretations only slightly, and they concluded that item interpretations were only moderately influenced by decisions related to using a weighting procedure.

All of the area indices proposed by Shepard et al. (1984) utilized discrete approximations and lacked sample standard errors to permit significant tests. Raju (1988, 1990) augmented these procedures by devising an index to measure continuous integration over unbounded intervals and derived standard errors permiting significant tests. Raju (1988) proposed setting the c parameter equal for both groups and estimating the signed area by

$$SA = (b_R - b_F) \, .$$

The unsigned area is estimated by

$$UA = \left| \frac{2(a_R - a_F)}{D a_R a_F} \ln(1 + \exp(\frac{D a_R a_F (b_R - b_F)}{a_R - a_F})) - (b_R - b_F) \right| \, .$$

Raju (1990) derived asymptotic standard error formulas for the signed and unsigned area measures that can be used to generate z tests to determine significance levels of DIF under conditions of normality.

Theoretically, Raju's procedure for measuring and testing the significance of the area between the ICCs for two groups is a significant advancement over procedures utilizing discrete intervals. Raju (1990) interpreted results derived from this procedure as sensible and found that the signed index and significant test provided results consistent with the MH procedure. Raju, Drasgow, and Slinde

(1993), analyzing data from a 45-item vocabulary trial test contrasting girls and boys and black and white students, found that the significance tests of the area measures identified the identical set of aberrant items as Lord's chi-square. Raju et al. (1993) set the alpha rate at 0.001 to control for Type I errors. Cohen and Kim (1993) found that in comparing Lord's chi-square to Raju's SA and UA, the two procedures produced similar results, although Lord's chi-square appeared slightly more powerful in identifying simulated DIF.

DIF as a Consequence of Multidimensionality

In all of the procedures thus far reviewed, researchers have either conditioned an item response on an observed test score or a latent ability estimate. Procedures using observed scores assumed that the total score has valid meaning in terms of the purported construct measured. The IRT procedures assumed responses to a set of items are unidimensional even though examinees' scores may reflect a composite of abilities. The potential for DIF can be conceptualized as occurring when a test consists of targeted ability, θ, and item responses are influenced by one or more nuisance determinants, η (Shealy & Stout, 1993a, 1993b). Under this circumstance, an item may be misinterpreted due to IRT model misspecification (Ackerman, 1992; Camilli, 1992; Oshima, 1989). If a misspecified, unidimensional IRT model is employed, the potential for DIF occurs if (a) the θ

means are not equal, (b) the η means are not equal, (c) the ratio $\sigma_\eta/\sigma_\theta$ are not equal, and (d) the correlations between the valid and nuisance dimensions are not equal (Ackerman, 1992).

The presence of multidimensionality in a set of items does not necessarily lead to DIF. For example, a quantitative ability test used to predict future college achievement may contain mathematical word problems requiring proficiency in reading skills. The test contains one primary dimension--quantitative ability; however, a second requisite measured skill--reading ability--is valid for the specific usage. A unidimensional analysis applied to such multidimensional data would weight the relative discriminations of the multiple traits to form a reference composite (Ackerman, 1992; Camilli, 1992). If the focal and reference groups share a common reference composite, DIF is not possible.

Since any test containing two or more items will to a degree be multidimensional, practitioners should define a validity sector to identify test items measuring approximately the same composite of abilities (Ackerman, 1992). In DIF studies, the conditioning variable should consist only of items measuring the same composite of abilities. If the dimensionality of the conditioning variable is not carefully defined, the DIF analysis is matching focal and reference group examinees on different composites of ability. This creates, in essence, the problem of trying to compare apples to oranges. The potential effect of this is to confound DIF with impact resulting in spurious interpretations (Camilli, 1992).

The effect of multidimensionality in DIF analyses has resulted in limited consistency across methods (Skaggs & Lissitz, 1992) and across differing definitions of the conditioning variable (Clauser, Mazor, & Hambleton, 1991). Further, R. L. Linn (1993) observed that rigorous DIF implementation to identify a proper set of test items may restrict validity. For example, on the SAT-Verbal (SAT-V), items with large biserial correlations to total score were more likely to be flagged than items with average or below average biserial correlations using MH. This finding suggested that traditional unidimensional DIF analyses, in part, might be statistical artifacts confounding group ability differences and item discrimination.

Differential item functioning procedures based upon a multidimensional perspective and conditioning on items clearly defined from a validity sector have the potential to reduce these problems (Ackerman, 1992). Further, a multidimensional approach should also facilitate DIF explanation (Camilli, 1992). Careful evaluation and conceptualization of multiple dimensions influencing item responses could aid in the identification and isolation of DIF’s causes.

SIBTEST. Shealy and Stout (1993a, 1993b) have formulated a DIF detection procedure within a multidimensional conceptualization. They conceptualize a test as measuring a unidimensional trait or reference composite--the target ability--that is influenced periodically by nuisance determinants. DIF is interpreted as the consequence of the differential effect of nuisance determinants functioning on an item or set of items.

The SIBTEST procedure employs factor analysis to identify a set of items that adheres to a defined validity sector. These items constitute the valid subtest, and the remaining items become the studied items. Examinees are divided into j strata based upon the valid subtest score, and the DIF index is estimated by

$$\beta_U = \sum_{j=1}^{J} p_j (\overline{Y_{Rj}} - \overline{Y_{Fj}}) \; ,$$

where p_j is the pooled weighting of focal and reference group examinees who achieve X = j. The value of β_U is identical to the value of STD P-DIF when the total number of examinees are the weighting group. Shealy and Stout (1993a) have referred to the standardization procedure as "progenitor" (p. 161) of SIBTEST. They present a significance test with the standard errror estimated by

$$SE(\beta)=\sqrt{\sum_{j=1}^{J} p_j^2\,[\frac{P_{Rj}(1-P_{Rj})}{N_{Rj}}+\frac{P_{Fj}(1-P_{Fj})}{N_{Fj}}]}.$$

With SIBTEST the total score on the valid subset serves as the conditioning criterion. The SIBTEST procedure resembles methods on which an observed test score is the criterion; although, it incorporates an adjustment to the item mean prior to comparing groups on these means. This adjustment is an attempt to remove that portion of group mean difference attributable to group mean differences on the valid targeted ability.

When the matching criterion is an observed score and the studied item is not included in the criterion score, group differences in target ability will tend to statistically inflate β. Consequently, SIBTEST employs a correctional procedure based upon regression and IRT theory. In effect, the purpose is to transform each observed mean group and ability level score, Y_{gj}, into a transformed mean so that the transformed score, $Y*_{gj}$, is a valid estimate of ability level score mean. This adjustment attempts to remove that portion of group mean differences that is attributable to group differences in the underlying targeted ability. Thus,

$$\overline{Y^*}_{Rj} - \overline{Y^*}_{Fj}$$

is an estimate of the difference in subtest true scores for the referenced and focal groups with examinees matched on ability levels. For this transformation to yield an unbiased estimate, the valid subtest must contain a minimum of 20 items (Shealy & Stout, 1993a).

SIBTEST is the only procedure based on conceptualizing DIF as a result of multidimensionality. Although it resembles the procedures that condition on observed scores, it offers a regression correction procedure that allows for conditioning on estimated true scores. Under simulated conditions it demonstrates good adherence to nominal error rates even when group target ability distribution differences are extreme, and it has been shown to be as powerful as MH in the detection of uniform DIF (Shealy & Stout, 1993a). Its multidimensional conceptualization potentially can lead to the identification of different nuisance determinants and greater understanding of DIF's causes (Camilli, 1992).

The major weaknesses of SIBTEST are its inability to assist the user in detecting nonuniform DIF and the need for 20 or more items to fit a unidimensional validity sector. With a relatively short test or subtest, this latter weakness would be problematic under some practical testing situations.

Methods Summary

After 20 years of development, a plethora of sophisticated DIF procedures have been devised. Each method approaches DIF identification from a fundamentally different perspective, and each method contains advantages and limitations. Currently, no consensus among DIF researchers exits regarding a single theoretical or practical best method. The design of this study reflected this lack of consensus. I selected five different procedures, each possessing theoretical or practical appeal, to assess item responses of examinees. The design of the study was not to compare the reliability and validity of the methods themselves, but to assess the similarity of results obtained from the methods when subpopulations were define in conceptually different ways.

Uncovering the Underlying Causes of DIF

The overwhelming majority of DIF researchers have focused on designing statistical procedures and evaluating their efficacy in detecting aberrant items. Few researchers have attempted to move beyond the methodological issues and examine DIF's causes. The researchers broaching this topic have experienced few successes and many frustrations.

Schmitt et al. (1993) proposed that explanatory DIF studies should begin with post hoc explorations of aberrant items and proceed to confirmatory experimental analyses. Researchers assessing DIF's causes use procedures that can be classified as (a) post hoc speculations, (b) hypothesis testing of item categories, (c) hypothesis testing using item manipulations, and (d) manipulation of other variables.

DIF can be attributed to a complex interaction between the item and the examinee (Scheuneman & Gerritz, 1990). Researchers are unlikely to find a single identifiable cause of DIF since it stems from both differences within examinees and item characteristics (Scheuneman, 1987). Researchers examining DIF from the perspective of examinee differences may uncover significant findings with implications for test-takers, educators, and policy makers. Scheuneman and Gerritz (1990) suggested that "prior learning, experience, and interest patterns between males and females and between Black and White examinees may be linked with DIF" (p. 129). Researchers examining DIF from the perspective of item characteristics may discover findings with strong implications for test developers and item writers. Test developers may need to balance content and item format to ensure fairness.

Post hoc evaluations, despite their limitations, dominate the literature (Freedle & Kostin, 1990; R. L. Linn & Harnisch, 1981; O'Neill & McPeek, 1993; Shepard et al., 1984; Skaggs & Lissitz, 1992). Speculations for causes of DIF begin with an interpretation of content similarities and patterns. For many researchers, the interpretation fails to go beyond a description of the observed item patterns

(O'Neill & McPeek, 1993; Shepard et al., 1984; Skaggs & Lissitz, 1992).

Hypothesis testing of item categories is a second, more sophisticated, means of uncovering explanations of DIF. Doolittle and Cleary (1987) and Harris and Carlton (1993) evaluated several DIF hypotheses on math test items. Doolittle and Cleary (1987) employed ACT Assessment Mathematics Usage Test (ACT-M) items and a pseudo-IRT detection procedure to analyze differences across item categories and test forms. Male examinees performed better on geometry and mathematical reasoning items, whereas female examinees performed better on computation items. Harris and Carlton (1993), using SAT-Mathematics (SAT-M) items and the MH procedure, concluded that male examinees did better on application problems and female examinees did better on more textbook-type problems.

Scheuneman (1987) analyzed 16 separate hypotheses concerning potential causes of DIF for black and white examinees by manipulating test items on the experimental portion of the GRE general test. The hypotheses, analyzed through log linear models, included examinee characteristics, such as test wiseness, and item characteristics, such as format. Complex interactions across groups, item pairs, and test content were observed.

Schmitt (1988) manipulated SAT-V items for white and Hispanic examinees to test four hypotheses derived from an earlier post hoc review. She employed the STDP-DIF index with ANOVA and found that Hispanic examinees were favored on antonym items that included a true cognate, a word with a common root in English and Spanish, and on reading passages containing material of interest to Hispanics. False cognates, words spelled similarly in both languages but containing different meanings, and homographs, words spelled alike in English but containing different meanings, tended to be more difficult for Hispanics. The differences were greater for Puerto Rican examinees, a group generally more dependent on Spanish, as compared to Mexican-American examinees.

K. K. Tatsuoka, R. L. Linn, M. M. Tatsuoka, and Yamamoto (1988) studied DIF on a 40-item fractions test. They initially analyzed examinees by dividing them into two groups based upon instructional methods. This procedure failed to provide an effective means of detecting DIF. However, upon subsequent review and analysis, they divided examinees into groups based upon solution strategies used in solving problems. With this grouping variable, they found DIF indices consistent with their a priori hypotheses. They concluded that the use of cognitive and instructional subgroup categories, although counter to traditional DIF research, contained potential for explaining DIF and diagnosing examinees' misunderstandings.

Miller and R. L. Linn (1988) considered the invariance of item parameters for the Second International Mathematics Study (SIMS) examination across different levels of mathematical instructional coverage. Although their principal concern was the multidimensionality of achievement test data as related to instructional differences and IRT model usefulness, they found that instructional differences could explain a significant portion of observed DIF. Using cluster analysis, they divided students into three instructional groups based upon teacher responses to an opportunity-to-learn questionnaire. The size of the differences in the ICCs for groups based upon instructional groups was much greater than differences observed in previously reported comparisons of black and white examinees. They interpreted these findings as supportive of R. L. Linn and Harnisch's (1981) postulation that what appears as item bias may in reality be "'instructional bias'" (p. 216).

Despite Miller and R. L. Linn's (1988) straightforward interpretation of instructional experiences, Doolittle (1984, 1985) found that instructional differences did not account for or parallel gender DIF on ACT-M items. He dichotomized high school math background into strong and weak, and compared a gender DIF analysis to a math background DIF analysis. In each analysis, approximately an equal number of items were detected; however, items that tended to favor female examinees did not favor low background examinees and vice versa. Correlations of DIF indices were negative, suggesting that gender DIF was unrelated to math background DIF.

Muthen, Kao, and Burstein (1991), analyzing the 40 core items of the SIMS test, found several items to be sensitive to instructional effects. In approaching DIF from an alternative methodological perspective, they employed linear structural modeling to assess the effects of instruction on latent mathematics ability and item performance. They found that instructional effects had negligible effects on math ability, but had significant influence on specific test items. Several items appeared particularly sensitive to instructional influences. They interpreted the identified items as less an indicator of general mathematics ability and more an indicator of exposure to a specified math content area.

In using linear structural modeling, Muthen et al. (1991) avoided the arbitrariness of defining group categories in a situation where group membership varied across items. The SIMS data permitted the estimation of instructional background for each of the 40 core items. Under most testing conditions, estimating examinee background differences to each item is impossible. However, general educational and psychological background variables can be modeled, and their relationship to unintended or nuisance dimensions estimated. Analyzing the relationship of theoretical causes of DIF to nuisance dimensions combines the approaches of Muthen et al. (1991) with Shealy and Stout (1993a, 1993b).

Summary

Researchers investigating the underlying causes of DIF have produced few significant results. After more than 10 years of DIF studies, conclusions of test wiseness (Scheuneman, 1987) or Hispanic tendencies on true and false cognates (Schmitt, 1988) must be interpreted as meager guidance for test developers and educators. These limited results can be explained by problems inherent in traditional DIF procedures (Skaggs & Lissitz, 1992; K. K. Tatsuoka et al., 1988). Indices derived using observed total scores as the conditioning variable have been observed to be confounded with item difficulty (Freedle & Kostin, 1990) and item discrimination (R. L. Linn, 1993; Masters, 1988). Indices derived from IRT models are conceptualized from an unidimensional perspective, yet DIF is a product of multidimensionality (Ackerman, 1992; Camilli, 1992). Consequently, DIF detection procedures have been criticized for a lack of reliability between methods and across samples (Hoover & Kolen, 1984; Skaggs & Lissitz, 1992).

The traditional conceptualization of dividing examinees by demographic characteristics limits DIF's explanatory potential (Skaggs & Lissitz, 1992; K. K. Tatsuoka et al., 1988). The uninterpretability of findings may be because group membership is only a weak surrogate for variables of greater psychological or educational significance. For example, demographic categories (e.g., women or blacks) lack any psychological or educational explanatory meaning. Moving beyond demographic subgroups to more meaningful categories would expedite understanding of DIF's causes (R. L. Linn, 1993; Schmitt & Dorans, 1990; Skaggs & Lissitz, 1992; K. K. Tatsuoka et al., 1988). Although this conceptualization has been advocated, it has been used sparingly. Doolittle (1984, 1985), Miller and R.L. Linn (1988), Muthen et al. (1991) and K. K. Tatsuoka et al. (1988) used this conception and appeared to have reached promising, if incompatible, interpretations. Future researchers need to apply alternative approaches to DIF analyses to achieve explanatory power. Approaches advocated by Muthen et al. (1991) and Shealy and Stout (1993a, 1993b) provide sound methods that potentially permit the modeling of differing influences on item responses.

Gender and Quantitative Aptitude

Educational and psychological researchers have been concerned with gender differences in scores on quantitative aptitude tests (Benbow, 1988; Benbow & Stanley, 1980; Friedman, 1989; Hyde, 1981; Maccoby & Jacklin, 1974), and their implications for career opportunities (M. C. Linn & Hyde, 1989). Mathematics has been termed the "critical filter" that prohibits many women from having access to high-paying and prestigious occupations (Sells, 1978). Although gender differences in quantitative ability interact with development, with elementary children demonstrating no difference or differences slightly favoring girls, by late adolescence and early adulthood, when college entrance examinations are taken and critical career decisions are made, slight differences appear favoring boys (Fennema & Sherman, 1977; Hyde, Fennema, & Lamon, 1990). In studies linking gender differences in quantitative test scores with the underrepresentation of women in prestigious technical careers, analyses should be limited to tests taken in late adolescence or early adulthood that significantly influence career decisions and opportunities.

Significant and Important Test Score Differences

Standardized achievement tests utilizing representative samples (e.g., National Assessment of Educational Progress, High School and Beyond) and college admissions tests utilizing self-selected samples (e.g., SAT, ACT, and GRE) have been analyzed to ascertain gender differences. Gender differences found in representative samples are systematically different from those found in self-selected samples (Feingold, 1992). Women appear less proficient, relative to men, in tests of self-selected samples of applicants as compared to representative samples. However, female students interested in technical professions must successfully matriculate through a process that relies heavily upon admissions test scores. Therefore, in studying quantitative differences with the primary concern related to career decisions and opportunities, self-selected admission test scores are the most germane measures for analysis.

M. C. Linn and Hyde (1989) concluded from meta-analytic studies (Friedman, 1989; Hyde et al., 1990) that "average quantitative gender differences have declined to essentially zero" (p.19), and differences in quantitative aptitude can no longer be used to justify the underrepresentation of women in technical professions. Feingold (1988) assessing gender differences in several cognitive measures on the Differential Aptitude Test (DAT) and the SAT concluded that gender differences are rapidly diminishing in all areas. The one exception to this finding was the SAT-M (Feingold, 1988). Although mean differences had either substantially diminished or vanished on DAT measures of numerical ability, abstract reasoning, space relations, and mechanical reasoning, during the past 30 years, SAT-M differences have remained relatively constant.

Despite the finding that gender differences are disappearing on many mathematical ability tests, on the major college entrance examinations gender differences remain large (Halpern, 1992). Over the past three decades, gender differences on the SAT-M have remained between 40 and 50 points. During this period, men averaged 46.5 points higher on the SAT-M than women (National Center for Education Statistics, 1993). This difference can also be stated in units of an effect size of 0.39 d (in which d represents the difference between the means divided by the pooled standard deviation).

The trends regarding gender differences on the ACT-M are similar. The ACT-M scale ranges from 1 to 39 points, and the mean difference favoring male examinees from 1978 to 1987 was 2.33 points or 0.33 d (National Center for Education Statistics, 1993). This score differential has been relatively consistent and provides no indication of disappearing.

The greatest disparity between men's and women's mean scores occurs on the GRE-Quantitative (GRE-Q). For the 1986-87 and 1987-88 testing years, U.S. male examinees averaged 86 and 80 points higher than U.S. female examinees (Educational Testing Service, 1991). Transformed into effect sizes, these differences are 0.67 d and 0.62 d, respectively. Gender mean score differences on the GRE-Q, in large part, reflect gender differences in choice of major field. Particularly in the case of graduate admissions tests, mean scores are confounded with gender differences in choice of undergraduate major. Analyzing GRE-Q data by intended field of study provides a more accurate comparison. For examinees intending to major in mathematics, the sciences, or engineering in 1986-1987, mean score differences favoring men were 37 and 18 points, respectively (d = .35 and .19). For examinees intending to major in the humanities and education in the same testing year, mean score differences favoring men were 44 and 37 points, respectively (d = .36 and .31). Averaging across 11 identified intended fields of study, mean score differences favoring men were 40 points (d = .35) (Educational Testing Service, 1991). Although data was available for only the 1986-87 testing years, mean score differences and effect sizes appear to indicate that U.S. male examinees tend to score higher than U.S. female examinees on the GRE-Q in a pattern consistent with the SAT-M and ACT-M.

Despite changes in the curriculum and text materials that depict both genders in less stereotypic manners (Sherman, 1983) and reductions in gender differences on many mathematics tests (Feingold, 1988), on college admissions quantitative tests gender differences are significant and appear not to be diminishing. Due to the importance of these tests regarding college admission decisions and the awarding of financial aid, the disparity in scores tends to reduce opportunities for women (Rosser, 1989).

Predictive Validity Evidence

Although mean scores on quantitative admission scores are higher for men than women, women tend to earn higher grades in high school and college (Kimball, 1989; Young, 1994). Test critics cited this paradox as principal evidence that admission tests are biased against women (Rosser, 1989). Defenders of the use of college admission tests argued that other relevant factors explain this phenomenon (McCornack & McLeod, 1988; Pallas & Alexander, 1983). They postulated that women tend to enroll in major fields where faculty tend to grade less rigorously (e.g., women are more likely to major in the humanities whereas men are more likely to major in the sciences). Investigators analyzing differential predictive validity of college admissions exams, therefore, must consider gender differences in course enrollment patterns.

McCornack and McLeod (1988) and R. Elliot and Strenta (1988) generally found that, when differential course taking patterns were considered, SAT-V and -M coupled with high school grades were not biased in predicting achievement for men and women. McCornack and McLeod (1988) considered performance in introductory level college courses at a state university and used SAT composites with high school grade point average. They found no predictive bias when analyzing data at the course level. R. Elliot and Strenta (1988) considered performance in various college-level courses at a private university and utilized SAT composites with scores from a college placement examination and high school rank. They also found no gender bias in prediction. However, they interpreted the SAT-M, when used in isolation, as underpredictive of women's college achievement. Both studies were flawed in that they combined various predictors and found no bias. Had they separately studied SAT-M and high school grades, they might have arrived at a different interpretation.

Bridgeman and Wendler (1991) and Wainer and Steinberg (1992) conducted more extensive studies and concluded that, for equivalent mathematics courses, the SAT-M tends to underpredict college performance for women. Bridgeman and Wendler (1991) studied the SAT-M as a predictor of college mathematics course performance at nine colleges and universities. They divided mathematics courses into three categories and found that, in algebra and pre-calculus courses, women's achievement was underpredicted and, in calculus courses, no underprediction occurred.

The most extensive study to date concerning the predictive validity of the SAT-M was conducted by Wainer and Steinberg (1992). Analyzing nearly 47,000 students at 51 colleges and universities, they concluded that, for students in the same relative course receiving the same letter grade, the SAT-M underpredicted women's achievement. Using a backward regression model, they estimated that women, earning the same grades in similar courses, tended to score roughly 25-30 points less on the SAT-M.

Summary

Researchers and test users have been troubled by the consistent findings than men tend to outperform women on quantitative admissions exams, although women generally outperform men in high school and college courses. The principal explanation offered for this paradox is gender differences in course taking. Researchers investigating the relationship of quantitative admission tests and subsequent achievement, controlling for course taking patterns and course performance, have concluded that, in equivalent mathematics courses, the tests underpredict women's achievement. Although the underprediction is not as large as mean score differences, quantitative admission tests do appear to be biased in underpredicting women's college achievement. It is recognized that predictive bias and DIF are fundamentally distinct; however, the determination of predictive bias in quantitative admission tests makes them an evocative instrument for analysis.

Potential Explanations of DIF

This study will approach DIF from the perspective of examinee characteristics. When analyzing DIF explanations from this perspective, theoretical explanations of predictive bias offer a reasonable point of departure. Kimball (1989) presented three theorectical explanations for the paradoxical relationship of gender differences on admissions test scores and college grades: (a) men have greater mathematical experience which enables them to more easily solve novel problems, (b) women tend to develop rote learning styles whereas men tend to develop autonomous learning styles, and (c) men tend to prefer novel tasks whereas women tend to prefer familiar tasks. To these three theorectical explanations, I would submit a fourth explanation related to test-taking behavior--differences between men in women in test anxiety.

Differences in Mathematics Background

It is well documented that as students enter high school and proceed toward graduation boys tend to take more mathematics courses than girls (Fennema & Sherman, 1977; Pallas & Alexander, 1983). During the 1980s, high school boys averaged 2.92 Carnegie units of mathematics whereas high school girls averaged 2.82 Carnegie units (National Center for Education Statistics, 1993). Although high school girls entered the upper-track ninth grade mathematics curriculum in slightly greater numbers than boys, by graduation, boys outnumbered girls in advanced courses such as calculus and trigonometry. High school boys were more likely to study computer science and physics than girls (National Center for Education Statistics, 1993). These trends continue as students enter college. During the 1980s, men slightly outnumbered women in achieving undergraduate mathematics degrees, and overwhelmingly outnumbered women in attaining undergraduate degrees in engineering, computer science, and physics. Gender disparities became even greater in the attainment of graduate degrees in mathematics, engineering, computer science, and physics (National Center for Education Statistics, 1993).

Researchers investigating the relationship between mathematics background and test scores have found that, when enrollment differences are controlled, gender differences on mathematical reasoning tests are reduced (Fennema & Sherman, 1977; Pallas & Alexander, 1983; Ethington & Wolfle, 1984). Gender score differences on the SAT-M, when high school course taking was controlled, were reduced approximately by two-thirds (Pallas & Alexander, 1983) and by one-third (Ethington & Wolfle, 1984).

These studies analyzed total score differences controlling for course background. Miller and R. L. Linn (1988) and Doolittle (1984, 1985) analyzed item differences controlling for instructional differences, but their results were contradictory. Background differences offer a plausible explanation for DIF that implores additional investigation.

Rote Versus Autonomous Learning Styles

Boys tend to develop a more autonomous learning style which facilitates performance on mathematics reasoning problems and girls tend to develop a rote learning style which facilitates classroom performance (Fennema & Petersen, 1985). Socialization patterns at home and in school tend to create these two distinct, gender-based, learning styles. Students displaying an autonomous learning style tend to do better, are more motivated, and are more likely to persevere on difficult tasks presented in a novel and independent format. Students displaying rote learning behavior tend to do well applying memorized algorithms learned in class and are heavily dependent upon teacher direction. Often, these students tend to choose less challenging tasks when given an option. This dichotomy is congruent with the finding that girls tend to perform better on computational problems and boys tend to perform better on application and reasoning problems (Doolittle & Cleary, 1988; Harris & Carlton, 1992).

The autonomous versus rote learning style theory is consistent with the literature addressing gender socialization patterns and standardized test performances. Before it can be further applied, however, it must be more completely operationalized (Kimball, 1989). To validate this theory, researchers must demonstrate that boys and girls approach the study of mathematics differently, and then relate learning styles to achievement on classroom assessments and standardized tests (Kimball, 1989).

Novelty Versus Familiarity

Kimball (1989) hypothesized that girls tend to be more motivated to do well and are more confident when working with familiar subject matter. Boys, on the other hand, tend to work harder and are more confident on novel tasks. Subsequently, girls tend to demonstrate higher achievement on familiar classroom assessments and boys tend to demonstrate higher achievement on novel standardized tests.

This theory is based on the work of Dweck and her colleagues (Dweck, 1986; E. S. Elliot & Dweck. 1987; Licht & Dweck, 1983) who related attributions to learning and achievement. Students with a performance orientation and low confidence tend to avoid difficult and threatening tasks. They prefer familiar, non-threatening tasks and seek to avoid failure. Students with a performance orientation and high confidence are more likely to select moderately challenging tasks. Consistent findings demonstrate that girls tend to have less confidence in their mathematical abilities than boys (Eccles, Adler, & Meece, 1984; Licht & Dweck, 1983). Girls are also more likely on standardized tests to leave items unanswered or mark "I don't know" when given this option (M. C. Linn, DeBenedictis, Delucchi, Harris, & Stage, 1987). Girls, more so than boys, attribute their success in mathematics to effort rather than ability and their failures to lack of ability (Fennema, 1985; Ryckman & Peckham, 1987). Therefore, due to less confidence in their abilities, girls generally are less motivated on novel mathematical tasks, find them more threatening, and perform less well.

Test Anxiety

Test anxiety has been hypothesized to adversely influence examinees' total scores on IQ tests, aptitude, and achievement tests. High test anxiety individuals tend to score lower than low test anxiety individuals of comparable ability (Hembree, 1988; Sarason, 1980). Because aptitude and achievement tests are not intended to include test anxiety as a component of total score, and because an estimated 10 million elementary and secondary pupils have substantial test anxiety (Hill & Wigfield, 1984), it exemplifies a nuisance factor influencing item responses.

Test anxiety has been theorized in both cognitive and behavioral terms (Hembree, 1988; Sarason, 1984; Spielberger, Gonzales, Taylor, Algaze, & Anton, 1978; Wine, 1980). Liebert and Morris (1967) proposed a two dimensional theory of test anxiety, consisting of _worry_ and _emotionality_. Worry includes any expression of concern about one's performance and consequences stemming from inadequate performance. Emotionality refers to the autonomic reactions to test situations (e.g., increased heartrate, stomach pains, and perspiration). Hembree (1988) used meta-analysis for 562 test anxiety studies and found that, although both dimensions related significantly to performance, worry was more strongly correlated to test scores. The mean correlations for worry and emotionality to aptitude/achievement tests were -0.31 and -0.15, respectively. Based upon a two dimensional model of test anxiety, Spielberger et al. (1978) proposed the _Test Anxiety Inventory_ (TAI).

Wine (1980) proposed a cognitive-attentional interpretation of test anxiety in which examinees who are high or low on test anxiety experience different thoughts when confronted by test situations. The low test anxious individual experiences relevant thoughts and attends to the task. The high test anxious individual experiences self-preoccupation and is absorbed in thoughts of failure. These task irrelevant cognitions not only create unpleasant experiences, but act as major distractions. Sarason (1984) proposed the Reactions to Test (RTT) scale based upon a cognitive, emotional, and behavioral model. The 40-item Likert-scaled questionnaire operationalized a four dimensional test anxiety model of (a) worry, (b) tension, (c) bodily symptoms, and (d) test-irrelevant thinking. Benson and Bandalos (1992), in a confirmatory cross-validation, found the four-factor structure of the RTT problematic. They speculated that misfit resulted from the large number of similarly worded items. Through a process of item deletion, they found substantial support for a 20-item four-factor model. To further validate the structure of test anxiety, Benson, Moulin-Julian, Schwarzer, Seipp, and El Zahhar (1991) combined the TAI and the RTT to formulate a new scale. The Revised Test Anxiety scale (RTA) was validated with multi-national samples and further refined (Benson & El Zahhar, 1994).

The cognitive and emotional structure of math anxiety is closely related to test anxiety. Richardson and Woolfolk (1980) demonstrated that math anxiety and test anxiety were highly related, and mathematical testing provided a superb context for studying test anxiety. They reported correlations between inventories of test anxiety and math anxiety ranging near 0.65. They commented that "(t)aking a mathematics test with a time limit under instructions to do as well as possible appears to be nearly as threatening as a real-life test for most mathematics-anxious individuals" (p. 271).

Children in first and second grade indicate inconsequential test anxiety levels, but by third grade test anxiety emerges and increases in severity until sixth grade. Female students at all age levels tend to possess higher test anxiety levels than male students at all grade levels (Everson, Millsap, & Rodriguez, 1991; Hembree, 1988). Some behavioral and cognitive-behavioral treatments have been demonstrated to effectively reduce test anxiety and lead to increases in performance (Hembree, 1988). This finding supports the causal direction of test anxiety producing lower performance and test anxiety's multidimensional structure.

The preponderance of research on test anxiety has focused on the relationship of test anxiety to total score performance. Harnisch & R. L. Linn (1981) speculated that in cases of model misfit, forces such as test anxiety might unduly influence performance at the item level. High test anxiety individuals may find some items differentially more difficult than other test items.

Summary

I have reviewed several different methods of identifying DIF. The MH, in large part because of its computational efficiency, has emerged as the most widely used method. It is limited in terms of its flexibility, and as researchers continue to search for underlying explanations of DIF, it limitations will become more apparent. Logistic regression models (Swaminathan & Rogers, 1990) provide an efficient method that has greater flexibility than MH and potentially models theoretical causes of DIF. Raju's (1988) IRT signed and unsigned area measures supply a theoretically sound method of contrasting item response patterns. Shealy and Stout's SIBTEST (1993a, 1993b) conceptualizes DIF as a multidimensional phenomenon and defines a validity sector as the conditioning variable. Its sound theoretical foundation coupled with it computational efficiency and explanatory potential makes it perhaps the most comprehensive DIF procedure. These five approaches were employed for the study. Linear structural modeling was used to factor analyze item responses and define a valid subset of test items. The five methods of DIF were applied before validation and incorporating the findings of the validation study. Thus, the significance of validation on the consistency of DIF estimation was considered.

Gender DIF on quantitative test items will serve as the context for this study. The context was taken because of the paradoxical finding that men tend to score higher on standardized tests of math reasoning, although women tend to achieve equivalent or higher course grades. Gender, a common categorical variable in DIF studies, will be supplement by dichotomizing examinees into substantial and weak mathematics background and high and low test anxiety.

This study is based on the premise that gender differences serve as a surrogate for differences in background and test anxiety. The two variables were selected in an effort to explain DIF in terms consistent with theoretical explanations of gender differences in mathematics test scores and course achievement. Mathematics background has been applied in other DIF studies with inconsistent interpretations. Test anxiety is of interest to both educators and cognitive psychologists and is highly related to performance. The study is an attempt to determine if the use of these variables serves to improve the consistency of DIF indices, detection methods, and aid in illuminating its causes.

CHAPTER 3
METHODOLOGY

The present study was designed to investigate the inter-method consistency of five separate differential item functioning (DIF) indices and associated statistical tests when defining subpopulations by educationally significant variables as well as the commonly used demographic variable of gender. The study was conducted in the context of college admission quantitative examinations and gender issues. The study was designed to evaluate the effect on DIF indices of defining subpopulations by gender, mathematics background, and test anxiety. Factor analytic procedures were used to define a structurally valid subtest of items. Following the identification of a valid subtest, the DIF analysis was repeated. The findings of the DIF analysis before validation were contrasted with the DIF analysis based on the valid subset. A description of examinees, instruments, and data analysis methods is presented in this chapter.

Examinees

The data pool to be analyzed consisted of test scores and item responses from 1263 undergraduate college students. The sample consisted of 754 women and 509 men. I solicited the help of various instructors in the colleges of education and business, and in most cases, students participated in the study during their class time. Of the total sample of examinees, 658 individuals were tested in classes of the college of education, 483 individuals were tested in classes in the college of business, and 122 individuals were tested at other sites on campus. Women and examinees with little mathematics background were the largest groups in the college of education classes, and men and examinees with substantial mathematics background were the largest groups in the college of business classes (see Table A.1 of Appendix A for examinee frequencies by test setting, gender, and mathematics background). The majority of students received class credit for participating. No remuneration was provided to any participant. All students had previously taken a college admissions examination, and some of the students (approximately 37 percent) had taken the Graduate Record Examination-Quantitative Test (GRE-Q).

Instruments

The operational definition of a collegiate-level quantitative aptitude test was a released form of the GRE-Q. Test anxiety was operationally defined by a widely used, standardized measure, the Revised Test Anxiety Scale (RTA). The mathematics background variable was measured using the dichotomous response to an item concerning whether or not students had completed a particular advanced mathematics class at the college level (i.e., calculus). In the following sections, a more detailed description of each of these instruments is presented accompanied by technical information that supports use of the particular instruments or item for the purpose of the study.

Released GRE-Q

Each examinee completed a released form of the GRE-Q. The 30-item test, supplied by Educational Testing Service (ETS), was a 30-minute timed examination. The sample test contained "many of the kinds of questions that are included in currently used forms" (ETS, 1993, p. 39) of the GRE-Q. The test was designed to measure basic mathematical skills and concepts required to solve problems in quantitative settings. It was divided into two sections. The format of the first section, quantitative comparison items, measured the ability to reason accurately in comparing the relative size of two quantities or to recognize when insufficient information had been provided to make such a comparison. The format of the second section, employing multiple choice items, assessed the ability to perform computations and manipulations of quantitative symbols and to solve word problems in applied or abstract contexts. The instructional background required to answer items was described as "arithmetic, algebra, geometry, and data analysis," and as "content areas usually studied in high school" (ETS, 1993, p. 18).

The internal consistency of the test for the 1263 participants was relatively good, KR-20 = 0.79. In a pilot study, the sample test correlations with the GRE-Q for 55 examinees and with the Scholastic Aptitude Test-Mathematics (SAT-M) for 58 examinees were 0.67 and 0.79, respectively. Thus, the scores on the released GRE-Q were similar to scores examinees earned on other college admissions quantitative examinations.

Revised Test Anxiety Scale (RTA)

The RTA scale (Benson, Moulin-Julian, Schwarzer, Seipp, & El-Zahhar; 1991) was formed by combining the theoretical framework of two recognized measures of test anxiety--the Test Anxiety Inventory (TAI) (Spielberger, Gonzales, Taylor, Algaze, and Anton, 1978) and the

Reactions to Tests (RTT)(Sarason, 1984). The TAI, based upon a two-factor theoretical conception of test anxiety--worry and emotionality (Liebert & Morris, 1967), contained 20 items. Sarason (1984) augmented this conceptualization with a four-factor model of test anxiety--worry, tension, bodily symptoms, and test irrelevant thinking.

To capture the best qualities of both scales, Benson et al. (1991) combined the instruments to form the RTA scale. They intended that the combined scale would capture Sarason's four proposed factors. From the original combined set of 60 items, using a sample of more than 800 college students from three countries, they eliminated items on the basis of items (a) not loading on a single factor, (b) having low item/factor correlations, and (c) having low reliability. They retained 18 items each loading on the intended factor and containing high item reliability. The bodily symptoms subscale, containing only 3 items, was problematic due to low internal reliability. Consequently, Benson and El-Zahhar (1994) further refined the RTA scale and developed a 20-item scale with four factors and relatively high subscale internal reliability (see Table 4). With a sample of 562 college students from two countries, randomly split into two samples, they cross-validated the RTA scale and found approximately equivalent item-factor loadings, factor correlations, and item uniquenesses. Descriptive statistics for each subscale of the RTA for Benson and El-Zahhar's (1994) American sample and this study's sample are reported in Table 4. The instrument was selected because evidence of its reliability and construct validity compared favorably with that of other leading test anxiety scales used with college students.

Table 4
Descriptive Statistics for the 20-item RTA Scale

Scales	Benson - El Zahhar American Sample N = 202	Study Sample N = 1263
Total Scale	38.31 10.40 .91	39.17 9.37 .89
Worry (6)	11.61 3.59 .81	12.03 3.50 .80
Tension (5)	12.81 3.85 .87	13.01 3.68 .84
Test Irrelevant Thinking (4)	6.61 2.53 .81	6.79 2.60 .83
Bodily Symptoms (5)	7.54 2.79 .76	7.35 2.55 .76

Note. Number of items per subscale is in parentheses. First entry in each column is the mean, second entry is the standard deviation, and the third entry is Cronbach's alpha.

Mathematics Background

Researchers have experienced problems selecting the best approach to measure subjects' mathematics background (Doolittle, 1984). Typically, methods for classifying subjects' background include (a) asking subjects to report the number of mathematics credits earned or semesters studied (Doolittle, 1984, 1985; Hacket & Betts, 1989; Pajares & Miller, 1994) or (b) asking subjects a series of questions related to specific courses studied (Chipman, Marshall, & Scott, 1991). Asking subjects questions concerning their course background implies that one or two "watershed" mathematics courses qualitatively capture subjects' instructional background. To decide which of these two options to employ in this study, I conducted a pilot study to ascertain whether measuring examinees' mathematics background by quantitatively counting mathematics credits earned or by qualitatively identifying a watershed mathematics course was more useful.

In a pilot study, 121 undergraduates were asked to answer the five questions posed by Chipman et al. (1991) and report the number of college credits earned in mathematics (see Appendix C for questions and the scoring scheme used with Chipman et al., 1991). Subjects were divided at the median into two groups and classified as possessing substantial or little mathematics background.

The subjects were then divided by using their responses to the single question about successful completion of a college calculus course. The two methods of dividing the 121 subjects into two background groups had an 84% agreement rate; however, correlations of these two predictors with performance on the GRE-Q and SAT-M indicated that the dichotomous calculus completion question was more valid for students in this study. The pattern of relationships between these tests, the calculus question, and the number of mathematics credits earned indicated that for these college students, calculus completion had a stronger relationship to the test scores (r = .50 - .51) than the number of mathematics credits earned (r = .08 - .40) (see Table 5).

In a continuation of the pilot study, 41 examinees reported they had successfully taken a college calculus

Table 5

Correlations of Calculus Completion, SAT-M, GRE-Q, and College Mathematics Credits

	SAT-M	GRE-Q	Credits
Calculus Completion	.51(58)	.50(55)	.49(141)
Total Credits	.08(58)	.40(55)	- -

Note. The number in parentheses represents the number of subjects each correlation is based upon.

course, and 100 examinees reported they had not successfully taken a college calculus course. The 41 examinees reporting successful completion of a college calculus course had earned an average of 13.3 college mathematics credits. The 100 students reporting they had not successfully completed a college calculus course had earned an average of 5.7 college mathematics credits. Therefore, for this sample there was substantial evidence that calculus courses serve as a watershed to other more advanced mathematics courses, and that completion of a calculus course could be used to differentiate students in terms of mathematics background.

Subsequently, mathematics background was operationalized by having each examinee answer the following question: "Have you successfully completed a college-level calculus course?" Examinees responding yes were classified as having a substantial background, and examinees responding no were classified as having little background. Utilizing examinee responses to the question of calculus completion was justified because of (a) the high degree of agreement between calculus completion and students' college course backgrounds, (b) the higher correlation of calculus completion to students' SAT-M and GRE-Q scores than total mathematics credits to students' SAT-M and GRE-Q scores, and (c) the need to dichotomous the sample by mathematics background in applying the DIF procedures.

Analysis

Testing Procedures and Subpopulation Definitions

Prior to taking the released GRE-Q, examinees answered the Differential Item Function Questionnaire (see Appendix B). It contained demographic questions and the RTA scale. Examinees provided information regarding their gender, mathematics background, and test anxiety. Examinees were classified as having substantial or little mathematics background by answering the question concerning completion of a college calculus course. Of the 1263 participants, 626 reported that they had completed a college calculus course and 637 reported that they had not completed a college calculus course. Frequency counts and percentages of mathematics background by gender are presented in Table 6. Men and women did not possess similar mathematics backgrounds. In the sample, 64% of the men reported completing a college calculus class, whereas 40% of the women reported completing a college calculus class.

High and low test anxious groups were formed in the following manner. Examinees scoring in approximately the highest 45 percent of the distribution on the RTA scale were defined as possessing high levels of test anxiety.

Table 6

Frequencies and Percentages for Gender and Mathematics Background

	Mathematics Background		Total
	Substantial	Little	
Women			
n	301	453	754
Pct.	23.8	35.9	59.7
Men			
n	325	184	509
Pct.	25.7	14.6	40.3
Total			
n	626	637	1263
Pct.	49.6	50.4	100

Examinees scoring in the middle 10 percent of the distribution were defined as possessing moderate levels of test anxiety. Examinees scoring in approximately the lowest 45 percent of the distribution were defined as possessing low levels of test anxiety. For the analysis, examinees classified as possessing moderate levels of test anxiety were deleted, and item responses of high test anxiety examinees were compared to item responses of low test anxiety examinees. Women tended to be classified as having high test anxiety at greater rates than men.

Following the completion of the questionnaire, examinees answered the 30-item GRE-Q. Examinees received a standard set of instructions and were told they had 30 minutes to complete the test. Examinees were requested to do their best, and following the test, if they desired, they could learn their results.

DIF Estimation

The five different methods for estimating DIF were Mantel-Haenszel (MH) (Holland & Thayer, 1988), Item Response Theory-Signed Area (IRT-SA) and Item Response Theory-Unsigned Area (IRT-UA) (Raju, 1988, 1990), Simultaneous Item Bias Test (SIBTEST) (Shealy & Stout, 1993b), and logistic regression (Swaminathan & Rogers, 1990). A distinction was made between *uniform* and *alternate* measures. Uniform and nonuniform methods estimate DIF in fundamentally different ways. If nonuniform DIF exits, the two approaches produce unique findings (Shepard, Camilli, & Williams, 1984). Consequently, the five methods were divided into two groups. Mantel-Haenszel, IRT-SA, and SIBTEST formed the uniform measures of DIF. Logistic Regression and IRT-UA, methods capable of detecting nonuniform DIF, coupled with MH, formed the alternate measures of DIF. Although MH was not designed to measure nonuniform DIF, test practitioners have used it extensively indicating that in actual testing circumstances they assume nonuniform DIF is either trivial or a statistical artifact. By examining the relationships between the DIF indices estimated by MH to those estimated by IRT-UA and logistic regression, researchers will be able to determine if important information is lost when only uniform methods are used.

Mantel-Haenszel indices and tests of significance were estimated using SIBTEST (Stout & Roussos, 1992). Item Response Theory signed and unsigned indices and tests of significance were estimated using PC-BILOG 3 (Mislevy & Bock, 1990) in combination with SAS 6.03 (SAS Institute, Inc., 1988). SIBTEST indices and tests of significance were estimated using SIBTEST (Stout & Roussos, 1992). Logistic regression indices and tests of significance were estimated through SAS 6.03 (SAS Institute Inc., 1988). Thus, each of the 30 test items was analyzed with three different subpopulation definitions and five different DIF procedures, producing for each item 15 distinct indices and significance tests.

Structural Validation

The structural component of construct validation concerned the extent to which items are combined into scores that reflect the underlying latent construct

(Messick, 1988). The structural component is appraised by analyzing the interrelationships of test items. The released GRE-Q was structurally validated through factor analysis of the matrix of tetrachoric coefficients for the 30-item test for a subsample of examinees. Initially, the sample of 1263 examinees was randomly split into two subsamples. The first subsample was used for the exploratory study, and the second subsample was used to cross-validate the findings derived from the exploratory analysis.

The tetrachoric coefficient matrix was generated with PRELIS (Joreskog & Sorbom, 1989a). Factor analytic models using an unweighted least squares solution through LISREL 7 (Joreskog & Sorbom, 1989b) were used to assess item dimensionality and potential nuisance determinants.

Research Design

Prior to validation, I assessed the consistency of the combination of five DIF methods and three subpopulation definitions. The inter-method consistency of DIF indices was assessed through a multitrait-multimethod (MTMM) matrix. The inter-method consistency of DIF significant tests was assessed by comparing percent-of-agreement rates between DIF methods when subpopulations are defined by gender, mathematics background, and test anxiety.

A subset of unidimensional items was identified by applying factor analytic procedures. Problematic items and items contaminated by nuisance determinants were identified. Following structural validation, the DIF analysis was repeated. Utilizing the combination of DIF methods and subpopulation definitions, DIF indices and significant tests were generated for the subset of items. The consistency of the DIF indices and associated inferential statistics was assessed. The findings assimilating validation were compared to the preceding findings to appraise the effect of structural validation on DIF analyses.

DIF Research Questions

Research questions one through four addressed the consistency of DIF indices through two MTMM matrices of correlation coefficients. Research questions one through four were first applied to the analysis of uniform DIF procedures and the MTMM matrix derived from these coefficients (see Table 1 on page 9). The same set of questions were then applied to the alternate DIF procedures and the MTMM matrix derived from these coefficients (see Table 2 on page 10).

The first question applied to the uniform DIF procedures focused on the convergent validity coefficients often termed the monotrait-heteromethod coefficients (e.g., the correlation of DIF indices when the subgroup or trait is gender and the methods are MH and IRT-SA). Were the convergent coefficients based upon the subpopulations of mathematics background and test anxiety greater than the convergent coefficients based upon gender subpopulations?

Specific statistical hypotheses were formulated to provide criteria for addressing the research questions. Let $\rho_{MI(G)}$ represent the correlation between the MH and IRT-SA DIF indices for the 30 items when examinee subpopulations are defined by gender. Let $\rho_{MS(G)}$ represent the correlation between the MH and SIBTEST indices for the 30 items when examinees are defined by gender. Let $\rho_{IS(G)}$ represent the correlation between the IRT-SA and SIBTEST indices for the 30 items when examinees are defined by gender. Comparable notation will represent examinee subpopulations defined by mathematics background (M) and test anxiety (TA). Three families of statistical tests each with two a priori hypotheses were defined to answer the first research question for the uniform methods. They were as follows:

H1a: $\rho_{MI(M)} > \rho_{MI(G)}$,

H1b: $\rho_{MI(TA)} > \rho_{MI(G)}$,

H2a: $\rho_{MS(M)} > \rho_{MS(G)}$,

H2b: $\rho_{MS(TA)} > \rho_{MS(G)}$,

H3a: $\rho_{IS(M)} > \rho_{IS(G)}$, and

H3b: $\rho_{IS(TA)} > \rho_{IS(G)}$.

The first question applied to the alternate DIF procedures also addressed the convergent or monotrait-heteromethod coefficients. Were the convergent coefficients based upon the subgroups of mathematics background and test anxiety greater than the convergent coefficients based upon gender subpopulations?

Similarly, for the alternate procedures let $\rho_{MI(G)}$ represent the correlation between the MH and IRT-UA DIF indices for the 30 items when examinee subpopulations are defined by gender. Let $\rho_{ML(G)}$ represent the correlation between the MH and logistic regression indices for the 30 items when examinees are defined by gender. Let $\rho_{IL(G)}$ represent the correlation between the IRT-UA and logistic regression indices for the 30 items when examinees are defined by gender. Comparable notation will represent examinee subpopulations defined by mathematics background (M) and test anxiety (TA). In a similar manner, three families of statistical tests each with two a priori hypotheses were defined to answer the first research question for the alternate methods. They were as follows:

H1a: $\rho_{MI(M)} > \rho_{MI(G)}$,

H1b: $\rho_{MI(TA)} > \rho_{MI(G)}$,

H2a: $\rho_{ML(M)} > \rho_{ML(G)}$,

H2b: $\rho_{ML(TA)} > \rho_{ML(G)}$,

H3a: $\rho_{IL(M)} > \rho_{IL(G)}$, and

H3b: $\rho_{IL(TA)} > \rho_{IL(G)}$.

The most efficient statistical test of two dependent correlations within a correlational matrix that do not share a common variable is

$$Z^{*} = \sqrt{N-3}\,[\frac{(z_{jk} - z_{hm})}{\sqrt{2-2\bar{s}_{jk,hm}}}],$$

where z_{jk} and z_{hm} are Fisher z transformations of values taken from the MTMM matrix, and $s_{jk,hm}$ is the asymptotic covariance of r_{jk} and r_{hm} (Steiger, 1980). This statistic has a z distribution and is easily interpreted.

Steiger's modified z* was combined with a Bonferroni-Holm procedure to control Type I errors. Using directional hypotheses, nominal Type I error rates were set within each family of hypotheses at .025 (.05/2) for the larger z*-value and .05 for the smaller z* value.

The second research question addressed whether the convergent coefficients (monotrait-heteromethod coefficients) were higher than the discriminant validity coefficients measuring different traits by identical methods. Campbell and Fiske (1959) maintained that when heterotrait-monomethod coefficients become larger than convergent coefficients a strong method effect is apparent. This criterion required each convergent coefficient to be higher than the four comparison coefficients of the corresponding triangular submatrices. The analysis of this question was applied to the uniform MTMM matrix and the alternate MTMM matrix.

The third research question focused on whether the convergent coefficients (monotrait-heteromethod coefficients) were higher than the discriminant validity coefficients measuring different traits by different methods. Convergent coefficients lower than heterotrait-heteromethod coefficients imply that agreement on a particular trait is not independent of agreement on other traits (Campbell & Fiske, 1959). This criterion required each convergent coefficient to be higher than the other four coefficients in the same row and column of the square submatrix. The analysis of this question was applied to the uniform MTMM matrix and the alternate MTMM matrix.

The fourth research question required the pattern between the three traits to be similar for the same and different methods. When the number of traits is small (e.g., three or four), this criterion is usually examined by inspection of the rank order of the correlations (Marsh, 1988). Fulfillment of this criterion provided evidence of true trait correlations independent of the method of assessment (Campbell & Fiske, 1959). Again, the analysis of this question was applied to both MTMM matrices.

The final research questions involved the consistency of DIF significance testing between methods when subpopulations are defined in different ways. Each of the five DIF methods included a statistical test to determine items exhibiting significant levels of DIF. Using the conventional alpha level of .05, the 30 items were classified as differentially functioning or non-differentially functioning in each of the 15 cases. Within each of the three ways of conceptualizing subpopulations, the percent-of-agreement in classifying items was determined between the three uniform methods and the three alternate methods.

Three percent-of-agreement rates for the uniform methods and three percent-of-agreement rates for the alternate methods were calculated for gender. Percent-of-agreement rates for the uniform and alternate methods also were calculated for the mathematics background and test anxiety DIF analyses. It was hypothesized that for both the uniform and alternate methods the percent of agreement rates between methods was higher for the mathematics background analysis and the test anxiety analysis as compared to the gender analysis.

Structural Validation Study

Since the purpose of the study was to investigate the structure of the released GRE-Q and to validate these relationships, the sample was randomly split into two subsamples. The first subsample consisted of 669 examinees--393 women and 276 men, and the second subsample consisted of 594 examinees--361 women and 233 men. The first subsample was used to investigate the dimensionality of the GRE-Q and to identify a subset of unidimensional items. The second subsample was used to cross-validate the findings derived from the exploratory study.

Linear structural equation modeling (LISREL) was implemented to assess item dimensionality. My objective was to identify a subset of unidimensionally valid items and problematic items for the GRE-Q. Problematic items were defined as items that possessed substantial loadings on more than a single factor or items that did not have an adequate loading on the dominant factor. An adequate loading was operationalized as 0.30 or greater.

A matrix of item tetrachoric coefficients initially was analyzed with a one-factor solution using an unweighted least squares procedure. To learn which items might be potentially problematic, I analyzed the item standardized estimates and residuals. For items seen as possibly problematic, I appraised their relationship to the remaining items. To evaluate the goodness-of-fit for the unidimensional model, I interpreted the Bentler-Bonett GFI (1980) and the Tucker-Lewis GFI (1973).

Based upon the findings derived from the unidimensional model, several alternate models were hypothesized. The alternate models contained multidimensional items representing nuisance. To assess the accuracy of the hypothesized multidimensional models, I evaluated goodness of fit indices, interfactor correlations, item standardized estimates, and item residuals. I classified items as unidimensional if, throughout the analyses, they maintained adequate loadings on the dominant factor and low loadings on other factors. Following the analysis of the hypothesized multidimensional models, I defined a subset of unidimensional items.

After the subset of unidimensional items was defined, the DIF analysis was repeated. DIF indices and significant tests were generated for the structurally valid subtest using the five methods and three subpopulation definitions. To assess the consistency of DIF indices, I generated a MTMM matrix for the uniform methods and a MTMM matrix for the alternate methods. I applied the four research questions evaluating DIF indices through the MTMM matrix. To assess the consistency of DIF significance test, I compared the percent-of-agreement rates between methods for gender, mathematics background, and test anxiety.

The findings from the DIF analysis following validation then were contrasted with the findings of the DIF analysis using the full test with no structural validation. The primary objective was to identify the differences between the two analyses and the influence of validation on DIF methods.

Summary

The MTMM matrices were used to investigate the relationship between three measures of uniform DIF and three alternate measures of DIF with subpopulations defined by gender, mathematics background, and test anxiety. A primary concern was to detect if conceptualizing DIF in terms of relevant educational and psychological variables improved the consistency of DIF methods. DIF significance tests were assessed by contrasting item classification percent-of-agreement rates within subpopulation definitions and between DIF methods.

A second concern of the study was to learn the influence of validation on the consistency of DIF methods. A structural validation procedure was conducted to identify unidimensional items and problematic items. Problematic items were defined as items possessing large loadings on more than a single factor or items having inadequate loadings on the dominant factor. Following validation, using the valid, unidimensional items, the DIF analysis was repeated to determine the consequence of validation on the results of the study.

CHAPTER 4
RESULTS AND DISCUSSION

In this chapter, I first provide descriptive statistics and frequency distributions of the principal variables. Second, I present and discuss findings of the data analyses relevant to each hypothesis. Last, I offer results of data analyses not directly related to the major questions but that were insightful and informed the theoretical implications in the study.

Descriptive Statistics

Table 7 presents the means and standard deviations of the released Graduate Record Examination-Quantitative (GRE-Q) and the Revised Test Anxiety Scale (RTA) for the total sample, men and women, and examinees possessing substantial and little background in mathematics. The released GRE-Q contained 30 dichotomously scored items. The RTA contained 20 items, each scored on a 4-point Likert scale.

The mean score on the released GRE-Q for men was 2.87 points higher than the mean score for women or stated in units of an effect size of 0.60 d (in which d represents the difference between the means divdied by the pooled

Table 7

Mean Scores of the Released GRE-Q and the Revised Test Anxiety Scale (RTA) by the Total Sample, Gender, and Mathematics Background

Variable	n	M	SD
Total Sample	1263		
GRE-Q		17.65	4.96
RTA		39.17	9.37
Women	754		
GRE-Q		16.49	4.61
RTA		40.28	9.59
Men	509		
GRE-Q		19.36	4.99
RTA		37.52	8.80
Substantial Math Bkd	626		
GRE-Q		19.58	4.62
RTA		38.43	9.39
Little Math Bkd	637		
GRE-Q		15.76	4.55
RTA		39.92	9.31

standard deviation). This finding was consistent with GRE published data (Educational Testing Service, 1993). The mean score on the released GRE-Q was 3.82 points higher for examinees possessing substantial mathematics background as compared to those possessing little mathematics background (d = .83). The mean score and standard deviation on the released GRE-Q for the 542 examinees classified as having low test anxiety were 18.97 and 4.92, respectively. The mean score and standard deviation on the released GRE-Q for the 558 examinees classified as having high test anxiety were 16.30 and 4.76, respectively. (Examinee score frequencies on the GRE-Q are presented in Table A.2 of Appendix A by gender, mathematics background, and test anxiety.)

The 30 items on the released GRE-Q had an average item difficulty of 0.59. Item biserial correlations ranged from 0.09 to 0.59 with a mean of 0.39. The mean biserial correlation for women was 0.36, and the mean biserial correlation for men was 0.42. Generally, the item biserial correlations were above 0.30, although four items had lower biserial correlations (r_{b2} = 0.23, r_{b6} = 0.28, r_{b10} = 0.23, and r_{b11} = 0.09). Item difficulties and biserial correlations are reported in Table A.3 of Appendix A.

The mean score on the RTA for women was 2.76 points higher than the mean score for men (d = .30). The mean score on the RTA for examinees with little mathematics background was 1.49 points higher than the mean score for examinees with substantial mathematics background (d = .16). Thus, women tended to score lower than men on the GRE-Q and tended to possess higher levels of test anxiety. Furthermore, examinees with substantial mathematics background tended to score considerably higher on the released GRE-Q than examinees with little background, although, they tended to possess only slightly less test anxiety.

The intercorrelations of the released GRE-Q, RTA, and mathematics background for the total sample, and for women and men are presented in Table 8. Performance on the released GRE-Q was negatively related to test anxiety and positively related to mathematics background, but the relationship of test anxiety and mathematics background, although statistically significant, was comparatively weaker. Across all three groups, the relationship between GRE-Q performance and mathematics background was the strongest.

Research Findings

It was of principal concern to learn whether defining subpopulations by relevant educational or psychological

Table 8

Intercorrelations of the Released GRE-Q, RTA, and Mathematics Background for the Total Sample, Women, and Men

Subscale	1	2	3
	Total Sample (n = 1263)		
1. GRE-Q	--	-.28	.39
2. RTA		--	-.08
3. Math Background			--
	Women (n = 754)		
1. GRE-Q	--	-.24	.35
2. RTA		--	-.06
3. Math Background			--
	Men (n = 509)		
1. GRE-Q	--	-.28	.34
2. RTA		--	-.04
3. Math Background			--

variables, rather than by gender, would yield results that were more consistent in magnitude across various DIF detection methods and more consistent in decisions regarding items that were classified as "biased." Secondarily, because theorists argue that DIF is a consequence of item multidimensionality, it was important to determine the effect of structural validation (i.e., unidimensionality) on the consistency of DIF estimation. I designed four unique contexts to examine this problem. In the first two contexts, I evaluated uniform and alternate DIF estimates prior to structural validation. In the second two contexts, I evaluated uniform and alternate DIF estimates applying the findings of the validation study.

Within each context, DIF estimation results were assessed through five research questions. A multitrait-multimethod (MTMM) matrix was employed to answer the first four questions. The observation of interest was the DIF index estimated for each item under a combination of subpopulation definition and DIF method. Trait effects were the three subpopulation definitions, and method effects were the five DIF procedures. DIF procedures were divided into uniform procedures and alternate procedures. A MTMM matrix was estimated for the uniform DIF procedures, and a MTMM matrix was estimated for the alternate DIF procedures. The fifth question involved the comparison of percent-of-agreement rates in detecting aberrant items for each method when subpopulations were defined in succession by gender, mathematics background, and test anxiety.

Following uniform and alternate DIF estimation under each trait and subpopulation definition, the dimensionality of the 30 items was evaluated through factor analysis. I attempted to create a more unidimensional test and identify potential nuisance dimensions that detracted from item validity. The goal was to devise a more unidimensional subtest and apply the five research questions incorporating the findings. Thus, the first facet of the study focused on the influence of using educationally relevant variables on the consistency of DIF estimation; and the second facet of the study evaluated the significance of structural validation on the consistency of DIF estimation.

DIF Analysis Prior to Validation

Uniform DIF procedures. For the three uniform DIF procedures, Mantel Haenszel (MH), Item Response Theory - Signed Area (IRT-SA), and SIBTEST, indices were estimated for subpopulations defined by gender, mathematics background, and test anxiety. The item DIF indices are reported in Tables A.4, A.5, and A.6 of Appendix A. The nine uniform indices for the 30 items were correlated and formed the MTMM matrix presented in Table 9.

Table 9

Multitrait-Multimethod Correlation Matrix: Uniform DIF Indices

	MH-D			IRT-SA			SIBTEST-b		
	A	B	C	A	B	C	A	B	C
I.MH-D									
A.Gender	--								
B.MathBkd	.10	--							
C.TA	-.18	-.20	--						
II.IRT-SA									
A.Gender	.80	.12	-.03	--					
B.MathBkd	.26	.72	-.14	.42	--				
C.TA	-.07	-.32	.59	.00	-.21	--			
III.SIBTEST-b									
A.Gender	.93	.10	-.18	.77	.27	-.09	--		
B.MathBkd	.14	.94	-.30	.13	.65	-.34	.15	--	
C.TA	-.11	-.30	.91	.00	-.21	.61	-.08	-.36	--

Note: Because the sign of the MH-D is diametrically reversed from SIBTEST-b and IRT-SA, the positive and negative signs for values of MH-D are reversed when used with the other two methods.

As I noted earlier, DIF theorists have speculated that conceptualizing subpopulations by relevant educational or psychological variables would enhance the consistency of procedures. For this reason, I hypothesized that the procedures would demonstrate greater consistency when subpopulations were defined by mathematics background and test anxiety than by gender. The first research question was that there would be stronger relationships evidenced by convergent validity coefficients based on mathematics background and test anxiety than by the corresponding convergent validity coefficients based on gender.

Convergent coefficients for MH and IRT-SA methods were 0.80, 0.72, and 0.59 with subpopulations defined by gender, mathematics background, and test anxiety, respectively. Convergent coefficients for MH and SIBTEST methods were 0.93, 0.94, and 0.91 with subpopulations defined by gender, mathematics background, and test anxiety, respectively. Convergent coefficients for the IRT-SA and SIBTEST methods were 0.77, 0.65, and 0.61 with subpopulations defined by gender, mathematics background, and test anxiety, respectively. Steiger's modified z* was use to test, within each pairwise method combination, whether the coefficients for mathematics background and test anxiety were higher than the corresponding coefficients for gender. None of the tests was significant. The test statistics are reported in Table A.7 of Appendix A. Therefore, within each of the three possible pairwise combinations of DIF estimation methods, defining subpopulations by mathematics background or test anxiety as compared to gender failed to produce more consistent DIF index estimation.

The second research question required that each convergent coefficient be higher than the four comparison coefficients of the corresponding triangular submatrices. All nine convergent coefficients were higher than their four comparison heterotrait-monomethod coefficients. For example, the convergent coefficient using MH and IRT-SA methods for the trait mathematics background was 0.72. The comparison heterotrait-monomethod coefficients were 0.10, -0.20, 0.42, and -0.21. Finding all nine convergent coefficients to be higher than the comparison coefficients indicated that uniform DIF indices exhibited minimal variance related to methods of DIF estimation.

The third research question required each convergent coefficient to be higher than the other four coefficients in the same row and column of the square submatrix. All nine convergent coefficients were higher than the other four coefficients in the same row and column of the square submatrix. For example, the convergent coefficient using

IRT-SA and SIBTEST methods for the trait test anxiety was 0.61. The four comparison heterotrait-heteromethod coefficients were -0.09, -0.34, 0.00, and -0.21. Finding all nine convergent coefficients to be higher than the comparison coefficients provided strong evidence of agreement on particular traits.

The fourth research question required the pattern between the three traits to be similar for the same and different methods. This question was answered by analyzing the rank order of the correlations in the heterotrait submatrix triangles. The heterotrait coefficients of gender and mathematics background ranked highest in all submatrix triangles (M = 0.19); the heterotrait coefficients of gender and test anxiety ranked second highest in all submatrices (M = -0.08); and the heterotrait coefficients of mathematics background and test anxiety ranked lowest in all submatrices (M = -0.26). The low values of the heterotrait coefficients indicated minimal method variance.

Utilizing the three subpopulation definitions and a MTMM matrix for analysis, the uniform DIF indices demonstrated good consistency and were minimally influenced by method variance. The full MH-SIBTEST submatrix illustrated the consistency of estimation and the limited method influence. Examining this submatrix indicated that all three convergent coefficients were high (0.93, 0.94, 0.91), and the heterotrait coefficients were low ranging from -0.30 to 0.14. Campbell and Fiske (1959) commented that often when researchers observe high convergent coefficients, the heterotrait coefficients are correspondingly large. They posited that such a finding indicates convergent validity coefficients inflated by method variance. In the MH-SIBTEST submatrix, however, the heterotrait coefficients were generally low and the high convergence coefficients indicated true agreement on defined traits. Although the other convergent coefficients are not as high as the ones estimated for the MH-SIBTEST submatrix, they are relatively high and indicate good consistency and little method variance. The analysis of the MTMM matrix of uniform DIF indices with the released GRE-Q supported the use of the methods with the subpopulation definitions.

The final research question for uniform methods was designed to assess the consistency of DIF significant tests between methods when subpopulations are defined respectively by gender, mathematics background, and test anxiety. Inferential statistics for the 30 items are reported in Tables A.8, A.9, and A.10 of Appendix A. With subpopulations defined by gender, MH chi-square identified 5 aberrant items, IRT-SA z-statistic identified 5 aberrant items, and SIBTEST z-statistic identified 6 aberrant items. With subpopulations defined by mathematics background, MH chi-square identified 8 aberrant items, IRT-SA z-statistic identified 8 aberrant items, and SIBTEST z-statistic identified 6 aberrant items. With subpopulations defined by test anxiety, each uniform procedure identified 2 items as aberrant. The item classification percent-of-agreement rates are presented in Table 10.

Table 10

Percent-of-Agreement Rates of Inferential Tests by Gender, Mathematics Background, and TA Between DIF Methods: 30-Item GRE-Q

Procedure Combination	Gender	Mathematics Background	TA
I. Uniform Methods			
MH - IRT-SA	80.0	73.3	96.6
MH - SIBTEST	96.7	86.7	100
IRT-SA - SIBTEST	83.3	73.3	96.6
II. Alternate Methods			
MH - IRT-UA	83.3	70.0	93.3
MH - Log Reg	76.7	70.0	90.0
IRT-UA - Log Reg	76.7	56.7	96.7

The highest percent-of-agreement rate occurred when subpopulations were defined by test anxiety; the next highest rate occurred when subpopulations were defined by gender; and the lowest rate occurred when subpopulations were defined by mathematics background. Not by chance, the rank ordering of percent-of-agreement rates was the reverse of the rank ordering of aberrant items detected by subpopulation definitions. The interpretation of percent-of-agreement rates in this context is confounded by the number of items detected within each subpopulation definition. Consequently, it is impossible to disentangle the findings and conclude that greater detection consistency occurred with subpopulations defined by test anxiety. An unanticipated finding of the study was that, when subpopulation were defined by mathematics background, all methods identified a larger number of differentially functioning items. This unanticipated finding will be addressed later in the chapter.

Alternate DIF procedures. For the three alternate DIF methods, MH, Item Response Theory-Unsigned Area (IRT-UA), and logistic regression, indices were estimated for subgroups defined by gender, mathematics background, and test anxiety. The item DIF indices are reported in Tables A.4, A.5, and A.6 of Appendix A. The nine alternate indices for the 30 items were correlated and formed the MTMM matrix presented in Table 11.

Table 11

Multitrait-Multimethod Correlation Matrix: Alternate DIF Indices

	MH-D			IRT-UA			Log Req		
	A	B	C	A	B	C	A	B	C
I.MH-D									
A.Gender	--								
B.MathBkd	.03	--							
C.TA	-.17	.11	--						
II.IRT-UA									
A.Gender	.55	.11	-.05	--					
B.MathBkd	-.05	.49	.12	.34	--				
C.TA	-.22	-.05	.32	-.01	.05	--			
III.Log Reg									
A.Gender	.20	.10	.21	.44	.21	.64	--		
B.MathBkd	-.04	.21	.20	.13	.41	.69	.86	--	
C.TA	-.17	-.09	.18	-.09	.05	.74	.79	.84	--

Note. The absolute value of the MH-D index for each of the 30 items was used in estimating the correlation coefficients for this table.

The first set of hypotheses for alternate procedures was that there would be a stronger relationship for convergent validity coefficients based on mathematics background and test anxiety than convergent validity coefficients based on gender. Convergent validity coefficients for MH and IRT-UA methods were 0.55, 0.49, and 0.32 withsubpopulations defined by gender, mathematics background, and test anxiety, respectively. Convergent validity coefficients for MH and logistic regression methods were 0.20, 0.21, and 0.18 with subpopulations defined by gender, mathematics background, and test anxiety, respectively. Convergent coefficients for IRT-UA and logistic regression methods were 0.44, 0.41, and 0.74 with subpopulations defined by gender, mathematics background, and test anxiety, respectively. Steiger's modified z* was used to test, within each pairwise method combination, whether the coefficients for mathematics background and test anxiety were higher than the corresponding coefficient for gender. None of the tests was significant. The test statistics are reported in Table A.7 of Appendix A. Thus, for each of the three possible pairwise combinations of the DIF estimation methods, defining subpopulations by mathematics background or test anxiety failed to improve the consistency of the alternate DIF indices.

The second research question required that each convergent coefficient be higher than the four comparison coefficients of the corresponding triangular submatrices. The coefficient pattern of the MTMM matrix did not meet this requirement. Convergent validity coefficients for all three traits measured by MH and logistic regression and by IRT-UA and logistic regression were problematic. For example, the coefficient estimated for the trait mathematics background using MH and logistic regression was 0.41 and was less than the heterotrait-monomethod coefficients of 0.86 and 0.84. The heterotrait-monomethod coefficients for logistic regression of 0.86, 0.79, and 0.84 were higher than any of the comparable convergent coefficients. The high heterotrait-monomethod coefficients indicated considerable variance related to logistic regression. However, this finding will change dramatically with the inclusion of the validation study.

The third research question required that each convergent validity coefficient be higher than the other four coefficients in the same row and column of the square submatrix. Using this criteria, only 5 of the 9 convergent coefficients were acceptable. This inadequate pattern of convergent coefficients indicated a lack of convergence of DIF indices related to specific subpopulation definitions.

The fourth research question required the pattern between the three traits to be similar for the same and different methods. The question was answered by analyzing the rank order of the correlations in the heterotrait submatrix triangles. Assessing the rank order of the correlations produced no discernible pattern. For example, discriminant coefficients related to the traits gender and test anxiety ranged in value from -0.22 to 0.79. The most stable discriminant coefficients were related to the traits gender and mathematics background, but their coefficients demonstrated an extreme range from -0.05 to 0.86. Consequently, the MTMM matrix for alternate DIF indices did not meet the requirements of the fourth research question.

The alternate MTMM matrix lacked the clear trait convergence that was apparent in the MTMM matrix of uniform DIF indices. The low convergent coefficients indicated poor consistency of DIF estimation. Variance related to methods was problematic particularly with logistic regression. A lack of agreement upon specific traits further indicated poor consistency within subpopulation definitions. The general pattern of the coefficients in the MTMM matrix for the alternate methods raised serious questions concerning their application when subpopulations are defined by gender, mathematics background, or test anxiety.

The final research question for the alternate methods was designed to assess the consistency of DIF significant tests between methods when subpopulations are defined by gender, mathematics background, and test anxiety. Inferential statistics for the 30 items are presented in Tables A.8, A.9, and A.10 of Appendix A. When subpopulations were defined by gender, MH chi-square identified five aberrant items, IRT-UA z-statistic identified five aberrant items, and logistic regression chi-square identified four items. When subpopulations were defined by mathematics background, MH chi-square identified 8 aberrant items, IRT-UA z-statistic identified 11 aberrant items, and logistic regression chi-square identified 12 aberrant items. When subpopulations were defined by test anxiety, MH chi-square procedure identified two aberrant items, IRT-UA z-statistic identified no aberrant items, and logistic regression chi-square identified one aberrant item. The item classification percent-of-agreement rates for the alternate methods are reported in Table 10.

The highest percent-of-agreement rate occurred when subpopulations were defined by test anxiety; the next highest rate occurred when subpopulations were defined by gender; the lowest rate occurred when subpopulations were defined by mathematics background. The rank ordering was the reverse as the number of aberrant items detected by subpopulation definition. This phenomenon was observed when percent-of-agreement rates were assessed for the uniform methods. Once again, analyzing the percent-of-agreement rates was confounded with the number of aberrant items detected within each subpopulation definition. For this reason, no accurate conclusions can be made concerning the differences of DIF estimation under different subpopulation definitions.

Test Validation and Dimensionality

Prior to conducting test validation, the total sample was randomly divided into two subsamples. The first subsample consisted of 669 examinees, and the second subsample consisted of 594 examinees. The first subsample provided information to explore the dimensionality of the test and define a valid subtest. The second subsample supplied information to cross-validate interpretations derived from the first subsample.

The exploratory study. I hoped to define through the exploratory study a subset of items that assessed the intended-to-be-measured dimension of the GRE-Q and potential nuisance determinants that hindered item validity. Theorists have posited that DIF is a consequence of one or more nuisance determinants interacting with the intended-to-be-measured dimension differentially for a defined subpopulation (Ackerman, 1992; Camilli, 1992; Shealy & Stout, 1993b). To initially study the dimensionality of the 30-item test, I factor analyzed the matrix of item tetrachoric coefficients. To assess the fit of various models and potential nuisance determinants, I analyzed goodness-of-fit indices, item standardized estimates, and item residuals.

A unidimensional model for the 30-item test indicated limited model fit (Bentler-Bonett GFI = 0.81, Tucker-Lewis GFI = 0.82). The unidimensional standardized estimates are reported in Table A.11 of Appendix A. Four problematic items (2, 6, 10, and 11) exhibited small loadings and large residuals. The tetrachoric coefficients of the four items to the remaining items were extremely small and, at times, negative. The low and negative coefficients of the problematic items impeded model fit and indicated that examinees responded to these items differently than to the remaining items. Furthermore, the interrelationships of the four items were low or negative indicating no single nuisance determinant. Table 12 presents the interrelationships of the four problematic items plus the unidimensional standardized estimates. Combinations of the four items were tried as

Table 12

Tetrachoric Correlations and Standardized Estimates for the Four Problematic Items: Exploratory Sample

	Item 2	Item 6	Item 10	Item 11
Item 2	(.15)			
Item 6	.126	(.24)		
Item 10	-.068	.115	(.23)	
Item 11	-.062	.170	.120	(.16)

Note. Numbers in the diagonal of the matrix are the unidimensional standardized estimates.

nuisance determinants, but none contained adequate model fit. For this reason, each item detracted independently from unidimensionality and was a self-contained source of unique nuisance variation.

I factor analyzed the remaining 26 items with a unidimensional model and assessed the goodness-of-fit indices, standardized estimates, and residuals to determine potential nuisance dimensions. The unidimensional model indicated improved fit over the total test (Bentler-Bonett GFI = 0.87, Tucker-Lewis GFI = 0.87). Standardized estimates for all items were greater than 0.30 and are reported in Table A.12 of Appendix A. Based upon item content, estimates, and residuals, various models with hypothesized nuisance determinants were factor analyzed, but none indicated better fit than the unidimensional model. Consequently, the most interpretable solution was that the GRE-Q contained 26 unidimensional valid items and 4 problematic items each being an independent source of nuisance variation.

Cross-validation. To examine the stability of the finding that the 4 items were problematic and the remaining 26 items formed a unidimensional valid test, a cross-validation was done using the second subsample. In the cross-validation, the same factor pattern was tested with the second subsample. The general item and factor patterns suggested by the exploratory study were specified, but item loadings and residuals were estimated without constraints. Initially, employing all 30 items, a unidimensional model was generated. Fit for this model was comparable to that for the exploratory sample (Bentler-Bonett GFI = .80, Tucker-Lewis GFI = .82). The standardized estimates are reported in Table A.11 of Appendix A. Although some item estimates changed in cross-validation, the estimates for three of the four problematic items remained low (< 0.30). Notwithstanding, the standardized estimate of Item 6 was 0.32.

The four problematic items from the exploratory study were deleted, and a unidimensional confirmatory factor analysis was generated for the 26 items with the cross validation sample. Model fit improved compared to the full test, but degenerated slightly in comparison to the exploratory study (Bentler-Bonett GFI = 0.84, Tucker-Lewis GFI = 0.85). The standardized estimates are reported in Table A.12 of Appendix A. The estimates ranged from 0.33 to 0.65. The general findings of the cross validation confirmed that items 2, 10, and 11 were problematic and impeded test interpretation. The remaining items had substantial loadings on the dominant factor. Item 6 was borderline problematic and of questionable validity.

The problematic items. Test items 2, 6, 10, and 11 are presented in Figure 1. To better understand these items, Figures 2, 3, 4, 5, 6, and 7 present plotted logistic regression curves (LRC). The total score for the 26 valid items was plotted on the horizontal axis, and the probability of a correct response was plotted on the vertical axis. LRCs were used instead of IRT item characteristic curves (ICC) because they are less restrictive than the two- or three-parameter ICCs. ICCs assume the lower asymptotes are zero or equal for both subpopulations, the upper asymptote is one, and the function of the ICC is monotonically increasing.

Directions: Each of the questions... consist of two quantities, one in column A and one in column B. You are to compare the two quantities and choose
A if the quanitity in Column A is greater;
B if the quanitiy in Column B is greater;
C if the two quanities are equal;
D if the relationship cannot be determined from the information given.

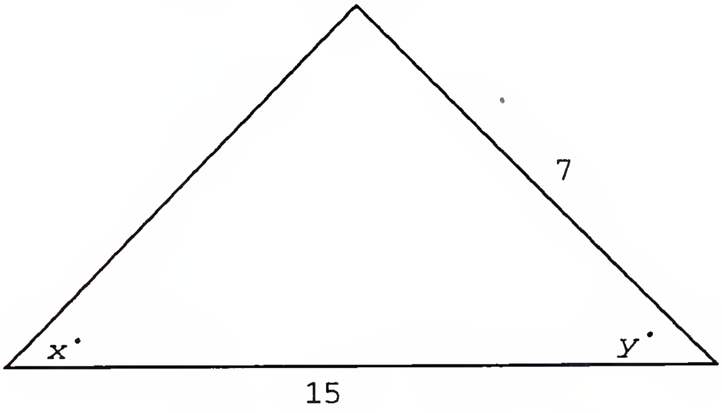
	Column A	Column B
2.	100.010-0.009	100.000+0.002

p is the probability that a certain event will occur, and $p(1-p) \neq 0$

6.	p	p^2

The average (arithmetic mean) of 2 positive integers is equal to 31 and each of the integers is greater than 26.

10.	The greater of the two integers	36



11.	x	y

Figure 1: The Four Problematic Test Questions.

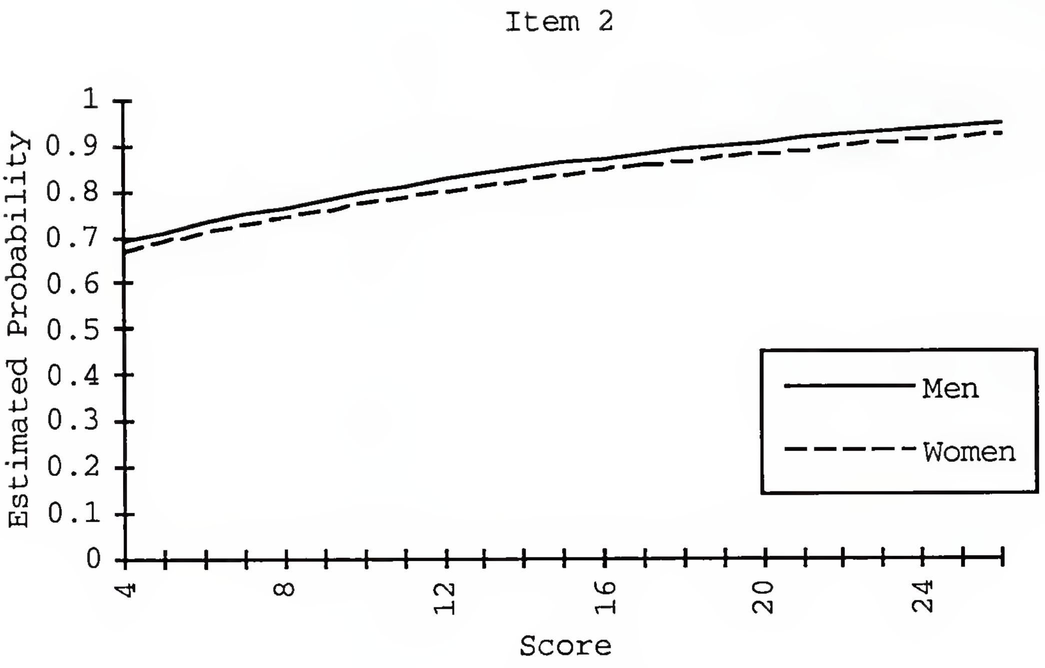


Figure 2: LRCs for Women and Men on Item 2.

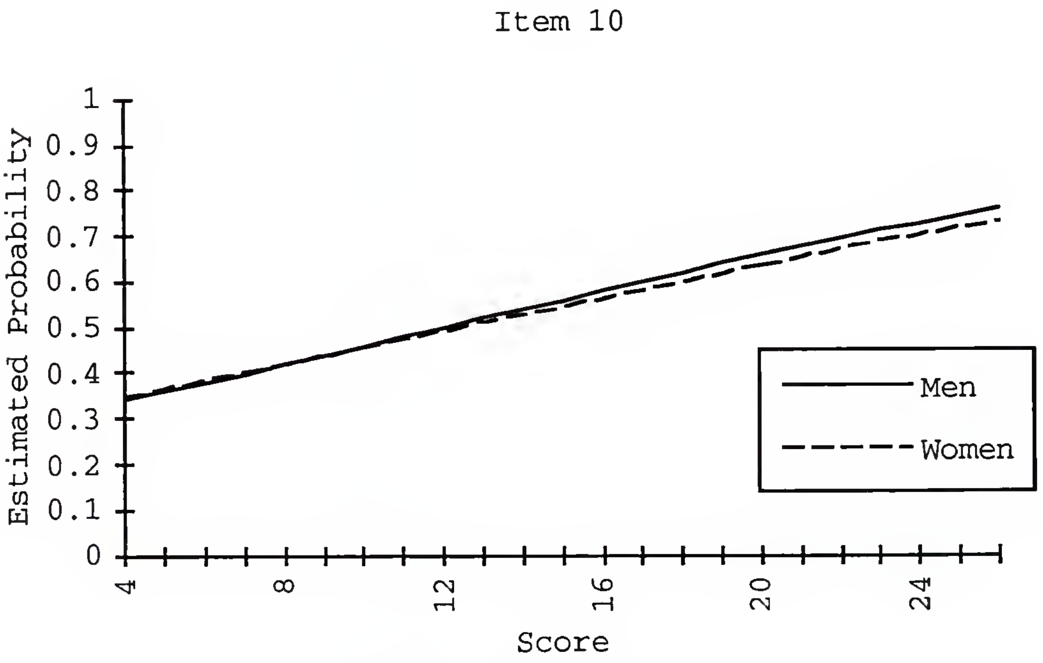


Figure 3: LRCs for Women and Men on Item 10.

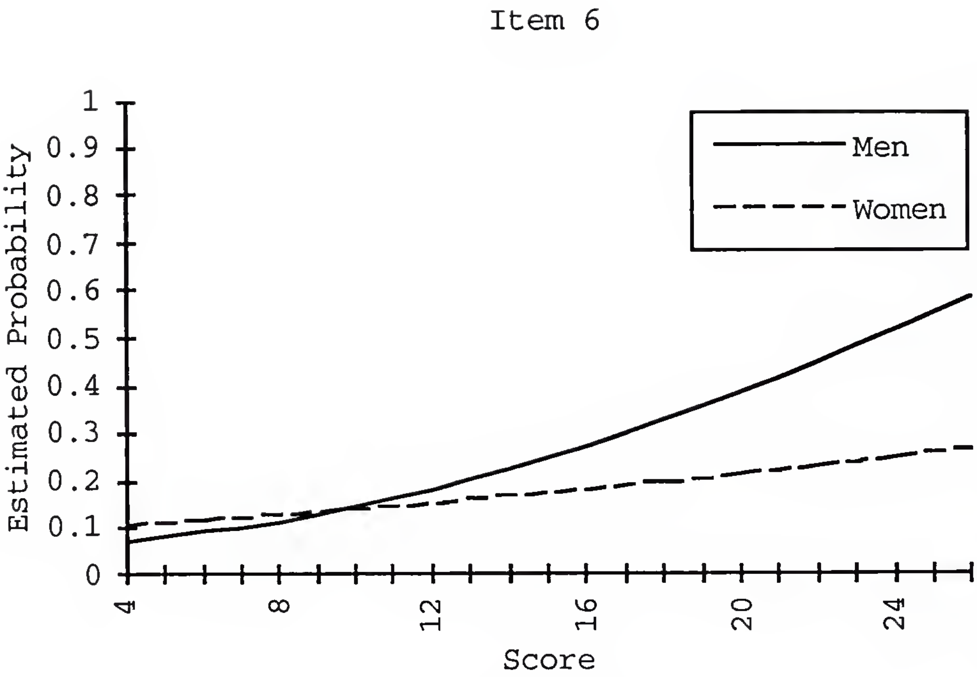


Figure 4: LRCs for Women and Men on Item 6.

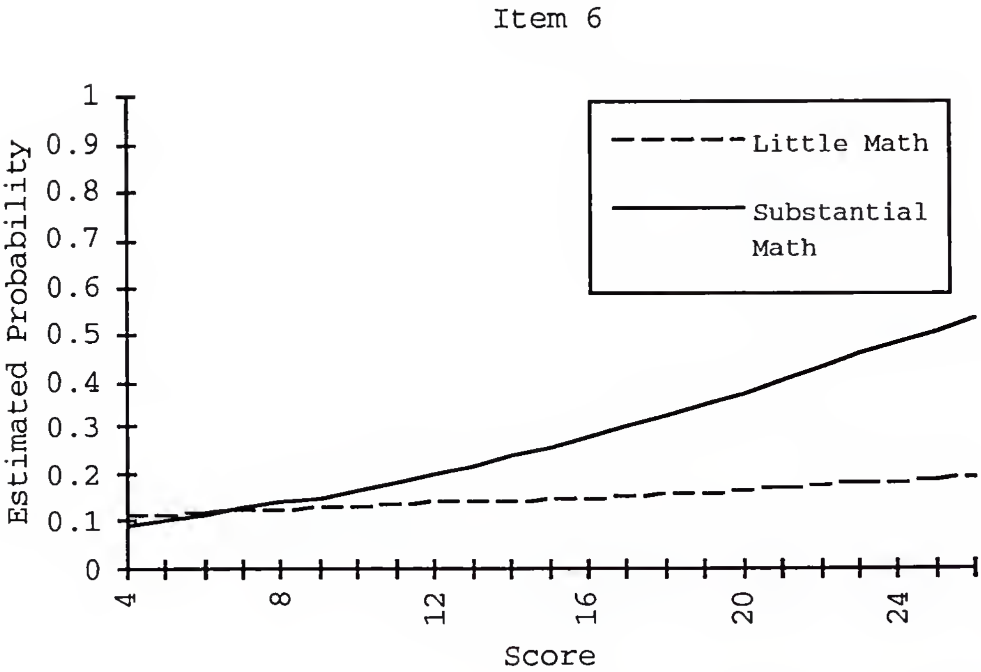


Figure 5: LRCs for Examinees with Substantial and Little Mathematics Background on Item 6.

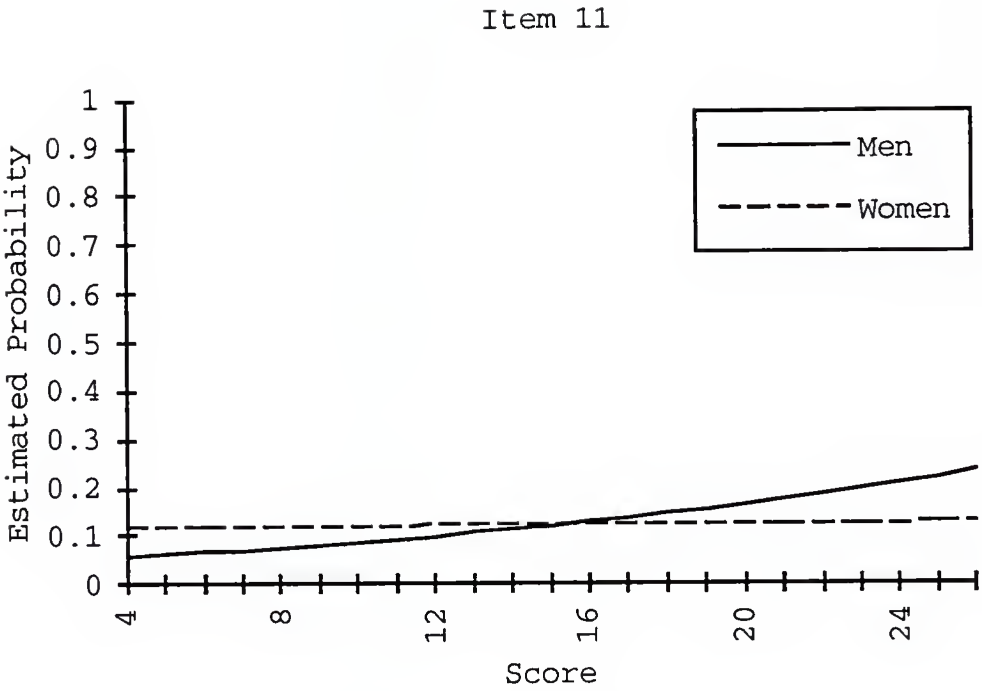


Figure 6: LRCs for Women and Men on Item 11.

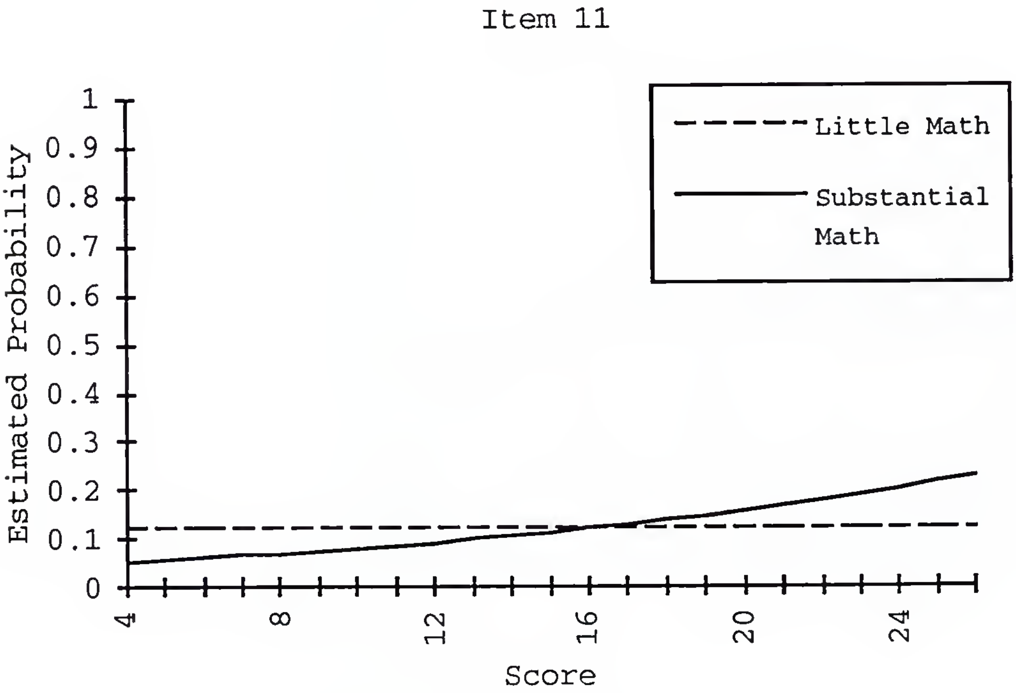


Figure 7: LCRs for Examinees with Substantial and Little Mathematics Background on Item 11.

Figures 2 and 3 present the LRCs for Item 2 and Item 10 by gender. Neither item was identified as containing significant levels of DIF, but both items had low discrimination as indicated by their low biserial correlations and plotted LRCs. Examinees of all abilities found Item 2 relatively easy (p = 0.85). Despite the difficulty of Item 10 being ideal (p = 0.57), as Figure 3 illustrates, examinees at the lowest ability level had a 35 percent probability of answering it correctly and examinees at the highest ability levels had less than a 70 percent probability of answering it correctly. Thus, Item 10 discriminated little between high ability and low ability examinees.

Figures 4 and 5 present the LRCs for Item 6 by gender and mathematics background. Conventional DIF analysis by gender and mathematics background found large and significant effects. The two figures are nearly identical and show little DIF for average to below average examinees, but, as ability increases, men and examinees with substantial mathematics background have higher likelihoods of correctly answering the item. The item is difficult (p = 0.24), however, it has questionable structural validity, is differentially functioning, and exemplifies gender DIF that can be explained by mathematics background. It should be noted, however, that in interpreting Figure 5 few examinees with substantial mathematics background scored 10 or less points on the GRE-Q. Consequently, the LRC for examinees with substantial mathematics background at low ability levels is estimated with extremely sparse data (see Table A.2 of Appendix A). From an educational perspective, students who have experience working with probabilities and performing mathematical operations on numbers between zero and one should be able to solve the item. Thus, specific curricular experiences should facilitate answering it. Men being more likely to enroll in mathematics and statistics courses possessed greater likelihoods, at above average ability levels, of correctly answering the item.

Figures 6 and 7 present the LRCs for Item 11 by gender and mathematics background. Item 11 had the highest difficulty and the lowest discrimination of all test items. High ability examinees were as unlikely to correctly answer it as low ability examinees. As Figure 7 illustrates, examinees of low ability with little mathematics background had slightly higher probabilities of answering it correctly than examinees of high ability with little mathematics background. Educationally, this item required knowledge of a geometric concept usually taught but not emphasized in high school geometry. Nearly all examinees failed to recall this concept and were forced to guess. Due to its content and statistical properties, Item 11 is a poor item. It does not correlate with other items on the test, and DIF methods fail to identify it as aberrant. Items so obscure that nearly all examinees are guessing lack any definable dimensionality and are problematic for conventional DIF analyses.

DIF Analysis Incorporating Structural Validation

In the follow-up DIF analysis, it was of principal interest to determine the influence of the structural validation study on DIF procedures and changes that might have occurred in the MTMM matrix and item detection. The five different DIF indices for the 26 unidimensional items with subpopulation definitions of gender, mathematics background, and test anxiety were generated. Utilizing the 26 items as the valid criterion, DIF indices were also estimated using MH, SIBTEST, and logistic regression for the 4 problematic items. DIF indices are reported in Tables A.13, A.14, A.15 of Appendix A. The estimated DIF indices for Item 11 using logistic regression were 110.45, 228.50, and 242.67 by gender, mathematics background, and test anxiety, respectively. The other methods did not produce such exceedingly high estimates for Item 11. (I will comment later in this chapter on logistic regression's unique values for Item 11). Because of the exceedingly high DIF index values with logistic regression for Item 11, in assimilating the findings of validation and assessing the consistency of DIF procedures, only the 26 unidimensional items were evaluated.

Uniform DIF procedures. Using the 26 items as the unit of analysis, a MTMM matrix of correlational coefficients was generated for the uniform DIF methods of MH, IRT-SA, and SIBTEST by the traits gender, mathematics background, and test anxiety. The MTMM matrix is presented in Table 13.

The pattern of coefficients using the 26-item test in the MTMM matrix for uniform DIF indices was similar to the matrix of the entire test. The answers to the specific research questions were generally the same. The first research question was that there would be a stronger relationship for convergent validity coefficients based on the subpopulation definitions of mathematics background and test anxiety than corresponding validity coefficients based on the subpopulation definition of gender. Convergent validity coefficients based on gender were equivalent to or slightly higher than convergent validity

Table 13

Multitrait-Multimethod Correlation Matrix of the Valid 26 Test Items: Uniform DIF Indices

	MH-D			IRT-SA			SIBTEST-b		
	A	B	C	A	B	C	A	B	C
I.MH-D									
A.Gender	--								
B.MathBkd	-.03	--							
C.TA	-.04	-.03	--						
II.IRT-SA									
A.Gender	.82	-.06	.08	--					
B.MathBkd	.19	.70	.13	.26	--				
C.TA	.13	-.27	.53	.11	-.19	--			
III.SIBTEST-b									
A.Gender	.90	.00	-.08	.72	.20	.02	--		
B.MathBkd	-.05	.95	-.12	-.06	.74	-.31	-.01	--	
C.TA	.01	-.10	.90	.08	-.14	.47	.00	-.15	--

Note. Because the sign of the MH-D is reversed from the SIBTEST-b and the IRT-SA, the positive or negative signs for values of the MH-D are reversed when used with the other two methods.

coefficients based on mathematics background and test anxiety. Consequently, following validation with subpopulations defined by mathematics background and test anxiety as compared to gender, greater consistency of DIF estimation was not achieved.

The convergent coefficients for the 26-item test were slightly lower than convergent coefficients for the entire test. The lower coefficients were most evident for the subpopulation group based on test anxiety. To a large degree, the lower observed coefficients were attributed to a reduction of variance in DIF indices when using the 26-item test.

The 26-item MTMM matrix for the uniform DIF indices successfully met the Campbell and Fiske (1959) guidelines for evaluating a MTMM matrix. All nine convergent coefficients were higher than their comparable heterotrait-monomethod coefficients. All nine convergent coefficients were higher than their comparable heterotrait-heteromethod coefficients. Like the MTMM matrix of uniform DIF indices based on the entire test, the 26-item MTMM matrix of uniform DIF indices indicated good convergence on the defined traits, minimal method variance, and supported the usage of DIF indices in the context of the subpopulation definitions.

The final research question was designed to assess the consistency of DIF significant tests. The inferential statistics are reported in Tables A.16, A.17, and A.18 of Appendix A. With subpopulations defined by gender, MH chi-square identified four aberrant items, IRT-SA z-statistic identified three aberrant items, and SIBTEST z-statistic identified five aberrant items. With subpopulations defined by mathematics background, MH chi-square identified seven aberrant items, IRT-SA z-statistic identified eight aberrant items, and SIBTEST z-statistic identified nine aberrant items. With subpopulations defined by test anxiety, MH chi-square and SIBTEST z-statistic both identified two items as aberrant, and IRT-SA z-statistic identified one item as aberrant. Table 14 presents the item classification percent-of-agreement rates. After, assimilating the structural validation findings, rates were approximately the same as the findings before validation.

Alternate DIF procedures. Utilizing the 26 items as the unit of analysis, a MTMM matrix of correlational coefficients was generated for the alternate DIF methods of MH, IRT-UA, and logistic regression by the traits gender, mathematics background, and test anxiety. The MTMM matrix is presented in Table 15.

Table 14

Percent-of-Agreement Rates of Inferential Tests by Gender, Mathematics Background, and Test Anxiety Between DIF Methods: 26-Item Valid Test

Procedure Combination	Gender	Mathematics Background	TA
I. Uniform Methods			
MH - IRT-SA	80.8	80.8	96.2
MH - SIBTEST	96.2	84.6	100
IRT-SA - SIBTEST	84.6	80.8	96.2
II. Alternate Methods			
MH - IRT-UA	80.8	84.6	96.2
MH - Log Reg	76.9	69.2	84.6
IRT-UA - Log Reg	96.2	76.9	88.5

Several significant differences were observed when the MTMM matrix utilizing the validation study was compared to the MTMM matrix estimated before structural validation presented in Table 11. The convergent coefficients for subpopulations defined by gender remained approximately equivalent to or greater than the convergent validity coefficients for subpopulations defined by mathematics background and test anxiety. Thus, the consistency of DIF estimation with the alternate

Table 15

Multitrait-Multimethod Correlation Matrix of the Valid 26 Test Items: Alternate DIF Indices

	MH-D			IRT-UA			LogReg		
	A	B	C	A	B	C	A	B	C
I.MH-D									
A.Gender	--								
B.MathBkd	-.14	--							
C.TA	-.31	-.04	--						
II.IRT-UA									
A.Gender	.53	-.15	-.16	--					
B.MathBkd	-.01	.52	-.32	.00	--				
C.TA	-.31	-.02	.09	-.25	-.20	--			
III.LogReg									
A.Gender	.52	-.22	-.19	.97	.03	-.21	--		
B.MathBkd	-.10	.51	-.28	.00	.95	-.14	-.03	--	
C.TA	-.25	.00	.03	-.04	-.05	.75	.01	.03	--

Note. The absolute value of the MH-D index for each of the 26 items was used in estimating the correlation coefficients.

procedures did not improve when subpopulations were defined by mathematics background or test anxiety. Although the answer to this question remained the same following structural validation, the answer to the other questions changed dramatically.

Eight of the nine convergent coefficients were higher than the comparable heterotrait-monomethod coefficients. Convergent coefficients higher than heterotrait-monomethod coefficients indicated an agreement upon traits and minimal method variance. The elimination of method variance attributed to logistic regression was a significant change following validation. The heterotrait-monomethod coefficients for logistic regression prior to validation were 0.86, 0.79, and 0.84. These high coefficients were reduced to near zero levels (-0.03, 0.01, and 0.03) using the valid 26-item test indicating no method variance. Thus, the method variance that was observed with the total test disappeared when the validation study was incorporated into the DIF analysis. This significant finding is couple with the extremely high convergent coefficients for IRT-UA and logistic regression (0.97, 0.95, and 0.75) indicating a high level of consistency between two methods.

The reasons for the dramatic changes in the convergent coefficients were related primarily to the elimination of Item 11 from the test and an increase the variance of DIF indices following structural validation. The elimination of Item 11 was significant because the logistic regression procedure consistently found it to be problematic while the other methods were incapable of detecting high levels of DIF under its unique conditions. Item 11 was the most difficult and least discriminating item on the GRE-Q ($p = 0.13$ and $r_b = 0.09$). When its LRC was plotted by gender and mathematics background subpopulations (Figures 6 and 7 on page 135), it contained an interaction between group membership and ability. Although the LRC for Item 11 indicated considerable nonuniform DIF, under its unique conditions, IRT-based methods were unable to identify it as problematic because of its restrictive model. For example using the 30-item test, DIF indices for Item 11 with subpopulations defined by mathematics background were 0.50 and 5.67 for IRT-UA and logistic regression, respectively. MH and SIBTEST were also inadequate because they were not designed to identify nonuniform DIF. Consequently, when DIF indices estimated through logistic regression were correlated with DIF indices estimated by the other methods, the inclusion of Item 11, a dramatic outlier, greatly reduced their magnitudes.

It is interesting to note that although the conditions of Item 11 produced problems in evaluating the magnitude of the convergent coefficients for IRT-UA and logistic regression, the conditions of Item 6 did not display a similar effect. Item 6 had similar but not as extreme item characteristics as Item 11 ($p = 0.24$ and $r_b = 0.28$). Its LRC also suggested an interaction between group membership and ability (Figures 4 and 5 on page 124). Despite these similarities, the nonuniform DIF indices for Item 6 were highly consistent. Therefore, I must emphasize that IRT model restrictions become troublesome only under the extreme conditions of high or low item difficulty values and exceedingly low item biserial correlations.

Incorporating the findings of the structural validation study and redefining the criterion measure produced greater variability among the estimated DIF indices. The greater variance of the DIF indices allowed for potentially higher coefficients. Thus, applying the findings of structural validation resulted in greater variance of DIF and potentially higher correlational coefficients.

Returning to the application of the Campbell and Fiske guidelines for evaluating a MTMM matrix, all nine convergent coefficients were now higher than the comparable heterotrait-heteromethod coefficients. This finding indicated strong convergence on specific traits. Generally, the matrix based upon the 26 unidimensional items met the Campbell and Fiske (1959) guidelines for evaluating a MTMM matrix. The revised matrix provided evidence of convergence on specified traits and minimal method variance.

The final question based on the 26-item valid test was designed to assess the consistency of DIF significant tests. The inferential statistics are reported in Tables A.16, A.17, and A.18 of Appendix A. With subpopulations defined by gender, MH chi-square identified four aberrant items, IRT-UA z-statistic identified three aberrant items, and logistic regression chi-square identified two aberrant items. With subpopulations defined by mathematics background, MH chi-square identified seven aberrant items, IRT-UA z-statistic identified nine aberrant items, and logistic regression chi-square identified nine aberrant items. With subpopulations defined by test anxiety, MH chi-square and logistic regression chi-square identified two aberrant items, and IRT-UA z-statistic identified one aberrant item. The item classification percent-of-agreement rates are reported in Table 14.

The percent-or-agreement rates following structural validation tended to increase. The one exception to this trend occurred with subpopulations defined by test anxiety. This findings counters the substantially higher rates observed with IRT-UA and logistic regression by gender and mathematics background. Under these conditions, the percentages increased from 76.7 to 96.2 and 56.7 to 76.9, respectively.

Two problems remained with the MTMM matrix for alternate methods despite the inclusion of structural validation. Low magnitudes of convergent coefficients based upon test anxiety raised questions about the trait's usefulness. Comparatively mediocre convergent coefficients for MH and IRT-UA and for MH and logistic regression illustrated the theoretical differences between uniform and nonuniform methods. The relatively low coefficients were an indication of the amount of nonuniform DIF present in the test data. These low coefficients demonstrated that strictly using uniform measures in this context would not be fully adequate. The findings regarding the nature of nonuniform DIF will be addressed under additional findings.

Additional Findings

Although not part of the focus of this investigation, other findings proved interesting. They are reported and discussed here because they serve to inform DIF theory and help clarify the interpretations of the study.

Subpopulation Differences in DIF Detection

Defining subpopulations by gender, mathematics background, and test anxiety affected the number of aberrant items detected. Regardless of DIF method, more aberrant items were detected defining subpopulations by mathematics background than by gender. Likewise, regardless of method, more aberrant items were detected defining subpopulations by gender than by test anxiety. When subpopulations were defined by test anxiety, with all methods, two or fewer items were detected as differentially functioning.

Because the validation study failed to define a nuisance determinant, it was impossible to fully account for why mathematics background led to a greater number of items detected. Perhaps, the consistent finding of mathematics background resulting in more items detected can be attributable to multidimensionality creating the potential for DIF. As I stated in Chapter 2, theorists have suggested that if a misspecified, unidimensional model is employed, the potential for DIF is present with any of the following four conditions (Ackerman, 1992; Camilli, 1992). They were as follows:

1. True ability means differ by group.

2. Nuisance ability means differ by group.

3. The ratios of the standard deviation for nuisance ability to the standard deviation for true ability differ by group.

4. The correlations of true ability and nuisance ability differ by group.

The first two conditions could have easily influenced the findings. The first condition of mean differences in true ability was apparent in that true ability mean differences were largest with subpopulations defined by mathematics background.

A second condition of nuisance ability mean differences could relate to subpopulation definitions and the nature of the study. On the GRE-Q, a test designed to measure quantitative aptitude, college mathematics instruction interacted with specific items. Specific instructional experiences assisted examinees in answering some items. I posited that instructional experience explained much of the observed gender DIF for Item 6. If item multidimensionality was related to instructional experiences, dividing examinees by mathematics backgrounds would manifest vastly different means on the nuisance determinant.

The finding of more aberrant items with subpopulations defined by mathematics backgrounds raises questions concerning the traditional DIF methods. Simple unidimensional mathematical models do not appear to represent the complexity of examinee responses to specific test questions adequately. Multidimensional models that incorporate demographic characteristics along with examinee background experiences, such as opportunity-to-learn, might enable researchers to better evaluate the fairness of specific items.

In regards to test anxiety and nuisance determinants, I must acknowledge the test had no real consequences for examinees. Consequently, it may not have provoked the intense levels of anxiety that normally are experienced by examinees taking college admissions tests. Because under these testing conditions high test anxiety examinees did not experience significantly higher levels of anxiety than low test anxious examinees, the mean differences on nuisance associated with test anxiety was slight. For this reason, the influence of test anxiety on performance was small, it interacted minimally with specific items, and appeared unrelated to nuisance.

Dimensionality and DIF

Although the validation study identified 26 items that were significantly associated with the dominant dimension of the test and four items that were problematic, it failed to identify items that were multidimensional. DIF theorists have proposed models of multidimensionality (Shealy & Stout, 1993a, 1993b), and they have discussed validity sectors and theoretical causes of multidimensionality (Ackerman, 1992; Kok, 1988). However, when applied to the test data, the analysis of standardized estimates, validity sectors, and residuals to identify multidimensional items was not productive.

In the context of the study, I was able to identify items that were highly associated with the dominant dimension of the test and items that were contained nuisance. Several multidimensional models were hypothesized and studied, and with the 26-item test, the interfactor correlations were 0.90 or greater. This finding indicated unidimensionality.

Nonuniform DIF

All DIF procedures are designed to assess uniform DIF, but only a few procedures are capable of detecting nonuniform DIF. Procedures assessing only uniform DIF imply that nonuniform DIF is either trivial or a statistical artifact. To study the efficacy of different methods, researchers simulate both uniform and nonuniform DIF. Swaminathan and Rogers (1990) in an attempt to demonstrate MH's inability to identify nonuniform DIF simulated data through a two-parameter IRT model. For items with nonuniform DIF, they set the difficulty parameter to zero and varied the discrimination parameter. In effect, they simulated nonuniform DIF with ICCs that crossed symmetrically. Such interactions created the worse case scenario where uniform methods have virtually no power. Although such simulations demonstrate conditions under which uniform procedures have little power, researchers must ask do symmetrical interactions occur with actual test data. In the context of this study, Item 8 with subpopulations defined by gender exhibited a symmetrical interaction. Item 8's LRC is presented in Figure 8. Although Item 8 represented the nonuniform pattern simulated by Swaminathan and Rogers (1990), the more common nonuniform DIF found in the study had LRCs that crossed at extreme ability levels. This more typical nonuniform DIF is illustrated in Figure 9 by

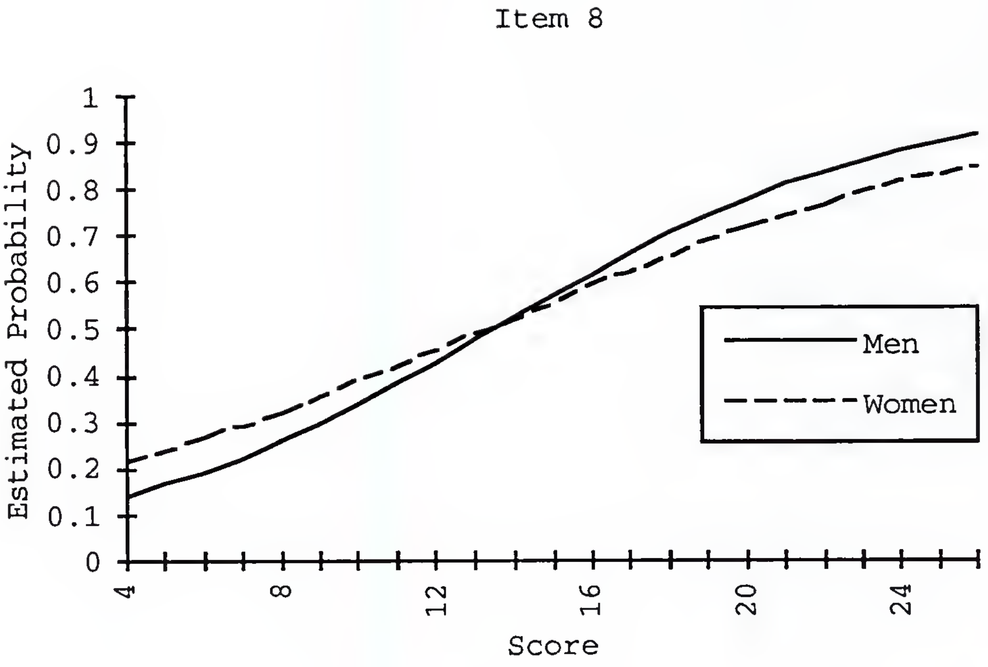


Figure 8: LRCs for Women and Men on Item 8 Illustrating the Symmetrical Nonuniform DIF Condition.

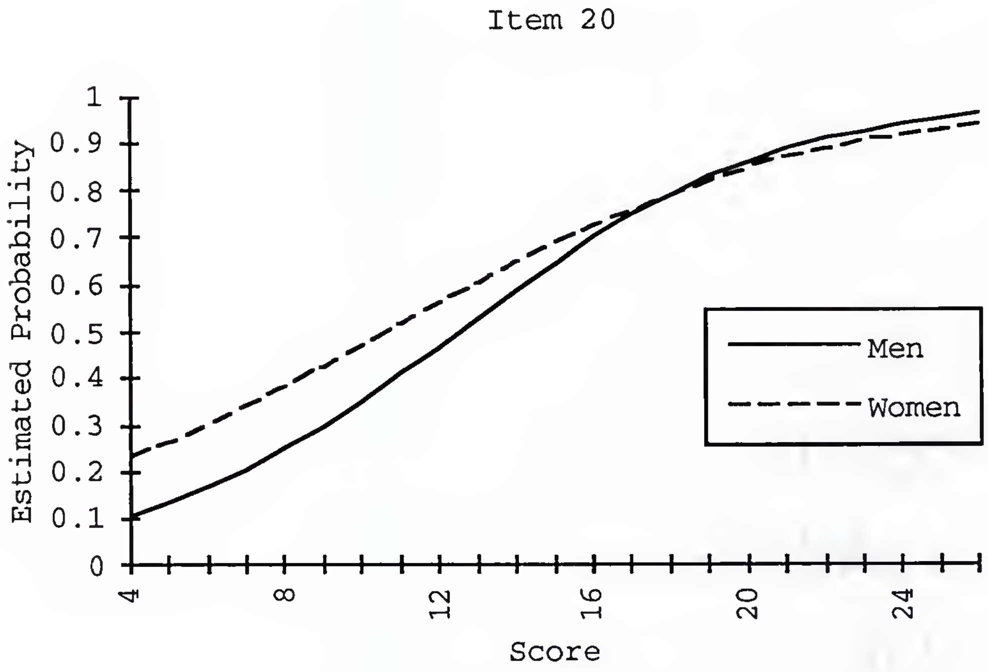


Figure 9: LRCs for Women and Men on Item 20 Illustrating the More Typical Nonuniform DIF Condition.

Item 20. Under this more typical case, the uniform methods possess considerably more power to detect DIF.

To further understand nonuniform DIF with actual data, IRT-SA indices were compared to IRT-UA indices. The unsigned area between the ICCs must be equal to or greater than the absolute value of the signed area. If the unsigned area is equal to the absolute value of the signed area, the two ICCs are completely uniform. The difference between the value of the unsigned area and the absolute value of the signed area can be interpreted as a crude indicator of the degree of nonuniformity. The mean absolute value of IRT-SA for the 30 items estimated under the three conditions was 0.29. The mean value of IRT-UA for the 30 items estimated under the three conditions was 0.41. Comparing these values indicated that a substantial amount of area between two ICCs was unaccounted for by IRT-SA. In the context of this study, nonuniform DIF did not occur as frequently as uniform DIF, and symmetrical patterns were very unusual. However, in assessing the test data, important information would be lost if only uniform measures were used.

CHAPTER 5
SUMMARY AND CONCLUSIONS

The two primary purposes of this study were to compare the consistency of five differential item functioning (DIF) methods (a) when subpopulations were defined by gender, mathematics background, and test anxiety, and (b) when the findings of a structural validation study were incorporated in the analysis. The study was conducted in the context of college admission quantitative examinations and gender issues. I chose this context because of the problematic predictive validity associated with mean differences between men and women on such tests. Within this context, findings would inform DIF researchers to the usefulness of defining subpopulations by psychological or educational variables and the effects of structural validation on DIF estimation.

The claim by Skaggs and Lissitz (1992) that defining subpopulations by psychologically or educationally relevant variables would yield more consistent DIF estimation was not substantiated; however, the findings indeed confirmed the importance of careful structural validation as a part of DIF studies (Ackerman, 1992; Camilli, 1992). Regarding the usefulness of defining subpopulations by psychological or educational variables, the attenuated coefficients of the multitrait-multimethod (MTMM) matrices suggested equivalent or better consistency when subpopulations were defined by gender as compared to mathematics background or test anxiety. The consistency of DIF indices with subpopulations defined by gender was approximately equivalent to the consistency of DIF indices with subpopulations defined by mathematics background. When the validation findings were incorporated, both modes of defining subpopulations were consistent. I had hypothesized that defining subpopulations by psychological or educational variables would result in higher consistency of results from various DIF estimation methods. In part, the failure to find significantly higher convergent validity coefficients with subpopulations defined by mathematics background was attributed to the reliability of measurement. The measurement of an examinee's gender possessed near-perfect reliability. Although the method for dividing examinees into two groups based on their mathematics background was validated, measurement reliability would not achieve the near-perfect reliability achieved in measuring gender.

Consequently, the convergent coefficients related to mathematics background were attenuated, whereas the coefficients related to gender were unattenuated.

The analysis of examinee responses to Item 6 provided an exemplary case of gender DIF that could be explained by mathematics background. With subpopulations defined by gender and mathematics background, the DIF indices for Item 6 were large and significant. Comparing the logistic regression curves (LRCs) of Figures 4 and 5 indicated that the response patterns for men and for examinees with substantial mathematics backgrounds were nearly interchangeable. Similarly, comparing the LRCs for women and examinees with little mathematics backgrounds suggested that their response patterns were nearly identical. For Item 6, the observed gender DIF was attributed to differences in mathematics background.

Unfortunately, only Item 6 contained such a straightforward interpretation. More frequently, items with significant gender DIF did not possess significant levels of DIF by mathematics background. As found in similar studies, the comparison of LRCs by gender and mathematics background was confusing and contradictory (Doolittle, 1985, 1986). Although the LRCs for Item 6 corroborated the positions of Miller and Linn (1988) and

Muthen, Kao, and Burstein (1991) in the use of educational background as a mode for investigating DIF, the results were mixed and unclear.

If the findings for studying DIF from the perspective of mathematics background produced conflicting interpretations, the findings for studying DIF from the perspective of test anxiety were straightforward. The consistency of DIF indices with subpopulations defined by gender was generally higher than the consistency of DIF indices with subpopulations defined by test anxiety. When structural validation findings were incorporated, each convergent coefficient by gender was higher than its corresponding coefficient by test anxiety. The comparatively low convergent coefficients by test anxiety had two probable causes. The first cause related to differences in measurement reliability for gender as opposed to test anxiety. Whereas gender was measured with near-perfect reliability, the Revised Test Anxiety Scale's estimated Cronbach's alpha was 0.89. The second and more detrimental cause was the lack of variance in DIF estimation with subpopulations defined by test anxiety. Item DIF indices for test anxiety subpopulations were extremely low, and they were rarely statistically significant. The low DIF indices suggested that, in the context of the study, defining subpopulations by test anxiety was of limited use for studying DIF. Because no actual consequences resulted from an examinee's test performance, participants did not feel levels of test anxiety comparable to those provoked by actual college admission tests. Consequently, many individuals who under true test conditions might have experienced extremely high levels of test anxiety, in the context of the study, felt only moderate or low levels of test anxiety. For these reasons, the consistency of DIF estimation by test anxiety was comparatively low, and test anxiety lacked explanatory power.

The consistency of DIF significant testing, as measured by item classification percent-of-agreement rates by gender, mathematics background, and test anxiety ranged from mediocre to perfect. For uniform DIF methods before validation, percent-of-agreement rates ranged from 73.3 to 100. For alternate DIF methods before validation, rates ranged from 56.7 to 96.7. For uniform DIF methods assimilating the findings of validation, rates ranged from 80.8 to 100. For alternate methods assimilating the findings of validation, rates ranged from 69.2 to 96.2. Under all method combinations, the rates were the highest with subpopulations defined by test anxiety. With subpopulations defined by gender, the rates were generally the second highest. With subpopulations defined by mathematics background, the rates were generally the lowest. The rank ordering of percent-of-agreement rates was the reverse of the rank ordering of aberrant items detected by subpopulation definitions. Therefore, it was impossible to disentangle these relationships and conclude that one subpopulation definition produced more consistent detection than another.

The finding that item classification percent-of-agreement rates for the study averaged in the mid-80s was better than the item classification rates observed by Hambleton and Rogers (1988). In comparing the MH to a method similar to IRT-UA, they found item classification rates near 80 percent. The finding of item classification agreement rates between 80 and 85 percent supports the position of Skaggs and Lissitz (1992) that DIF detection methods are mediocre at best. To illustrate, for a 30-item test using two methods each identifying 5 aberrant items, the worst-case scenario of no agreement on any aberrant items results in a rate of 66.7 percent. Statistical methods that improve this worst-case scenario rate to 80 or 85 percent have achieved limited improvement. Current DIF significance tests appear method dependent. For this reason, the findings supported researchers who advocate the importance of interpreting both the DIF index and significance test in evaluating the fairness of test items (Angoff, 1993; Burton & Burton, 1993; Dorans & Holland, 1993; Uttaro & Millsap, 1994).

The most prominent finding of the investigation was the effect of structural validation on DIF estimation. Before validation, the MTMM matrix of uniform DIF indices met the validity guidelines of Campbell and Fiske (1959). Subsequently, I found that the three uniform methods possessed high consistency and minimal variance related to methods. However, I observed that when the guidelines of Campbell and Fiske were applied to the MTMM matrix of alternate DIF indices severe problems appeared. The three alternate methods appeared to have low consistency and high levels of variance related to methods.

Factor analysis was employed to estimate item dimensionality and define a valid subset of items. In the exploratory study, 26 items were identified as having strong relationships to the dominant trait and 4 items were defined as problematic. Cross-validation generally supported this interpretation. Therefore, I deleted the 4 problematic items from the study and reevaluated DIF estimation.

The MTMM matrix of uniform DIF indices for the 26-item subtest was similar to the matrix before validation. The MTMM matrix met the validity guidelines of Campbell and Fiske (1959) indicating good consistency and minimal variance related to methods.

The MTMM matrix for alternate methods changed dramatically after employing the validation findings. Variance related to methods as identified by the heterotrait-monomethod coefficients was reduced to near zero. The convergent validity coefficients between the IRT-UA and logistic regression methods went from 0.44, 0.41, and 0.74 by gender, mathematics background, and test anxiety, respectively, to 0.97, 0.95, and 0.75 for the same subpopulation definitions. Theoretically, IRT-UA and logistic regression are highly related. Nevertheless, findings indicated that if a researcher ignored test validation and passively utilized the full test, DIF indices using IRT-UA or logistic regression possibly would be inconsistent and lead to poor decision-making.

Regarding the dramatic improvement of the MTMM matrix for alternate methods, it should be noted that the procedures I utilized differed from those employed by Ackerman (1992). Ackerman (1992) utilized structural validation to define a subset of valid items and then reestimated DIF indices for all items. He found for the valid subset generally low levels of DIF and concluded that DIF tended to disappear from the valid subset of items after validation. I utilized structural validation and defined a unidimensional subset of items, but reestimated DIF indices for only the valid items. Dropping the problematic items from the study resulted in more consistent nonuniform DIF indices. If I had maintained the problematic items in the analysis, the convergent coefficients between IRT-UA and logistic regression would have remained high and difficult to explain. This was a result of the erratic DIF estimates for the problematic items. For items unrelated to other items on a test, DIF indices appear to lack consistency.

The findings that defining subpopulations by mathematics background led to the detection of more aberrant items along with the erratic DIF indices found after validation raises questions concerning traditional DIF methods. Traditional DIF methods that are based on simple unidimensional models with subpopulations defined by demographic characteristics tend to result in inadequate model fit with actual test data. The simplistic models do not represent the complex interactions between examinees and test items. Examinee responses to test items are influenced by differences in educational backgrounds, opportunities-to-learn, and other life experiences. Traditional DIF models that consider only one trait or background differences are likely to be inadequate representations of actual test data. Future DIF research should incorporate multidimensional models similar to those developed by Reckase (1985, 1986) and include operationalized variables similar to the educational background variables defined by Miller and Linn (1988) and Muthen et al. (1991).

The findings further informed researchers concerning the need to measure and interpret nonuniform DIF. Numerous DIF studies have functioned from the perspective that nonuniform DIF is trivial or a statistical artifact (Burton & Burton, 1993; Freedle & Kostin, 1990; Harris & Carlton, 1993; Holland & Thayer, 1987; Scheuneman & Gerritz, 1990; Schmitt, 1988; Schmitt & Dorans, 1990; Zwick & Ercikan, 1989). Nonetheless, other researchers have simulated symmetrical interactions to demonstrate the need for nonuniform methods (Swaminathan & Rogers, 1990). In this study, examining the responses of examinees to the 30-item test, nonuniform DIF did not occur as frequently as uniform DIF and symmetrical nonuniform DIF occurred only once (Item 8 with subpopulations defined by gender).

Despite this observation, more subtle cases of nonuniform DIF occurred throughout the analysis and ignoring these could result in poor item analysis.

In analyzing items displaying nonuniform DIF, LRCs tended to intersect at either high or low levels of ability. In attempting to make valid decisions concerning the fairness of items, researchers need to evaluate possible nonuniform DIF in relationship to subpopulation ability distributions. Depending upon the differences in the distribution of subpopulations, ignoring nonuniform DIF could result in the inclusion of items unfair to a specific group. Recent efforts by Li and Stout (1994, 1995) to extend the SIBTEST to measure nonuniform DIF are significant advances in DIF theory and item validation.

Finally, a simplistic interpretation of the study's findings might be that uniform methods are easier to use, result in higher levels of consistency, do not require extensive validation, and are preferable to the nonuniform methods. The comparison of nonuniform IRT-SA and IRT-UA along with the plotting of the LRCs suggested that significant information would be lost if only uniform methods were employed. A thorough analysis of the findings confirmed the importance of careful test validation before DIF estimation, the significance of interpreting nonuniform DIF, and the need to interpret both the DIF index and significance test before eliminating items. The findings as well suggested that unidimensional DIF models are inadequate representations of the complex interactions between examinees and test items. This study further indicated that using mathematics background to define subpopulations and explain gender DIF has potential usefulness, although, in the context of this study, the use of test anxiety to define subpopulations and explain DIF was ineffectual.

APPENDIX A
SUMMARY STATISTICAL TABLES

Table A.1

Examinee Frequencies and Percentages for Gender and Mathematics Background by Test Location

	Mathematics Background		Total
	Substantial	Little	
	College of Education Classes (n=658)		
Women			
n	147	346	493
Pct.	22.3	52.6	74.9
Men			
n	74	91	165
Pct.	11.2	13.8	25.1
Total			
n	221	437	658
Pct.	33.6	66.4	100
	College of Business Classes (n=483)		
Women			
n	128	76	204
Pct.	26.5	15.7	42.2
Men			
n	191	88	279
Pct.	39.5	18.2	57.8
Total			
n	319	164	483
Pct.	66.0	34.0	100

Table A.1 -- continued.

	Mathematics Background Substantial	 Little	Total
	Other Test Locations (n=122)		
Women			
n	26	31	57
Pct.	21.3	25.4	46.7
Men			
n	60	5	65
Pct.	49.2	4.1	53.3
Total			
n	86	36	122
Pct.	70.5	29.5	100

Table A.2

GRE-Q Score Frequency Distributions by Gender, Mathematics Background, and Test Anxiety

Score	Freq	Percent	Score	Freq	Percent
		Women (n=754)			
3	0	0.0	17	62	8.2
4	0	0.0	18	42	5.6
5	3	0.4	19	61	8.1
6	8	1.1	20	48	6.4
7	7	0.9	21	45	6.0
8	14	1.9	22	34	4.5
9	20	2.7	23	22	2.9
10	26	3.4	24	19	2.5
11	23	3.1	25	10	1.3
12	45	6.0	26	13	1.7
13	53	7.0	27	5	0.7
14	54	7.2	28	2	0.3
15	64	8.5	29	1	0.1
16	72	9.5	30	1	0.1

Score	Freq	Percent	Score	Freq	Percent
		Men (n=509)			
3	1	0.2	17	42	8.3
4	1	0.2	18	28	5.5
5	0	0.0	19	36	7.1
6	2	0.4	20	34	6.7
7	4	0.8	21	43	8.4
8	4	0.8	22	44	8.6
9	5	1.0	23	35	6.9
10	11	2.2	24	34	6.7
11	5	1.0	25	24	4.7
12	17	3.3	26	21	4.1
13	20	3.9	27	14	2.8
14	17	3.3	28	10	2.0
15	21	4.1	29	6	1.2
16	29	5.7	30	1	0.2

Table A.2 -- continued.

Score	Freq	Percent	Score	Freq	Percent
Substantial Math Background (n=626)					
3	0	0.0	17	47	7.5
4	0	0.0	18	35	5.6
5	0	0.0	19	49	7.8
6	0	0.0	20	53	8.5
7	3	0.5	21	58	9.3
8	5	0.8	22	52	8.3
9	7	1.1	23	42	6.7
10	8	1.3	24	40	6.4
11	2	0.3	25	27	4.3
12	23	3.7	26	30	4.8
13	16	2.6	27	15	2.4
14	29	4.6	28	11	1.8
15	31	5.0	29	6	1.0
16	35	5.6	30	2	0.3

Score	Freq	Percent	Score	Freq	Percent
Little Math Background (n=637)					
3	1	0.2	17	57	8.9
4	1	0.2	18	35	5.5
5	3	0.5	19	48	7.5
6	10	1.6	20	29	4.6
7	8	1.3	21	30	4.7
8	13	2.0	22	26	4.1
9	18	2.8	23	15	2.4
10	29	4.6	24	13	2.0
11	26	4.1	25	7	1.1
12	39	6.1	26	4	0.6
13	57	8.9	27	4	0.6
14	42	6.6	28	1	0.2
15	54	8.5	29	1	0.2
16	66	10.4	30	0	0.3

Table A.2 -- continued.

Score	Freq	Percent	Score	Freq	Percent
Low Test Anxiety (n=542)					
3	0	0.0	17	36	6.6
4	0	0.0	18	28	5.2
5	0	0.0	19	41	7.6
6	2	0.4	20	48	8.9
7	5	0.9	21	42	7.7
8	4	0.7	22	42	7.7
9	5	0.9	23	26	4.8
10	9	1.7	24	31	5.7
11	13	2.4	25	23	4.2
12	22	4.1	26	24	4.4
13	19	3.5	27	12	2.2
14	28	5.2	28	10	1.8
15	25	4.6	29	5	0.9
16	40	7.4	30	2	0.4

Score	Freq	Percent	Score	Freq	Percent
High Test Anxiety (n=558)					
3	1	0.2	17	52	9.3
4	1	0.2	18	29	5.2
5	3	0.5	19	44	7.9
6	8	1.4	20	23	4.1
7	4	0.7	21	36	6.5
8	11	2.0	22	26	4.7
9	20	3.6	23	24	4.3
10	24	4.3	24	16	2.9
11	11	2.0	25	8	1.4
12	38	6.8	26	6	1.1
13	41	7.3	27	3	0.5
14	35	6.3	28	1	0.1
15	43	7.7	29	1	0.1
16	49	8.8	30	0	0.0

Table A.3

Released GRE-Q Item Difficulty Values (p) and Biserial Correlations (r_b) for the Total Sample, Women, and Men

	Total Sample		Women		Men	
Item	*p*	r_b	*p*	r_b	*p*	r_b
1	.91	.40	.92	.35	.90	.51
2	.85	.23	.83	.24	.88	.23
3	.83	.32	.79	.39	.88	.21
4	.46	.40	.41	.40	.53	.39
5	.68	.38	.65	.36	.72	.45
6	.24	.28	.18	.14	.32	.34
7	.76	.33	.74	.30	.81	.33
8	.59	.30	.55	.21	.66	.38
9	.46	.45	.38	.42	.58	.41
10	.57	.23	.54	.24	.61	.23
11	.13	.09	.12	.05	.15	.16
12	.51	.41	.45	.31	.60	.47
13	.56	.44	.53	.43	.60	.49
14	.69	.50	.66	.47	.74	.55
15	.44	.55	.38	.54	.53	.54

Table A.3 -- continued.

	Total Sample		Women		Men	
Item	p	r_b	p	r_b	p	r_b
16	.84	.59	.81	.57	.89	.57
17	.80	.52	.77	.52	.85	.51
18	.81	.43	.78	.40	.84	.46
19	.76	.44	.72	.45	.83	.40
20	.69	.39	.66	.32	.72	.50
21	.77	.37	.73	.36	.83	.33
22	.64	.40	.58	.34	.72	.47
23	.34	.38	.32	.35	.37	.42
24	.63	.56	.59	.53	.70	.60
25	.85	.50	.83	.46	.88	.62
26	.29	.31	.25	.21	.36	.34
27	.55	.40	.50	.33	.62	.45
28	.46	.42	.41	.34	.53	.46
29	.23	.36	.19	.31	.29	.35
30	.32	.41	.26	.33	.41	.44

Table A.4

DIF Indices for the 30 Items of the GRE-Q with Population Groups Defined by Gender

Item	MH-D	SIBTEST-b	IRT-SA	IRT-UA	LogReg
1	1.76	-.062	-0.80	0.80	0.84
2	-0.23	.020	0.22	0.28	0.23
3	-0.71	.050	1.41	1.60	1.72
4	0.08	-.016	-0.15	0.32	0.07
5	0.31	-.021	-0.28	0.28	0.32
6	-1.02	.081	1.05	1.21	1.98
7	-0.03	.007	0.07	0.25	0.15
8	-0.17	.024	-0.15	0.48	0.75
9	-0.89	.088	0.15	0.30	0.37
10	0.22	-.013	-0.16	0.47	0.13
11	0.15	.002	-0.04	0.28	3.92
12	-0.22	.022	0.08	0.31	0.39
13	0.74	-.065	-0.36	0.36	0.28
14	0.51	-.043	-0.27	0.27	0.18
15	0.14	-.019	-0.21	0.31	0.07

Table A.4 -- continued.

Item	MH-D	SIBTEST-b	IRT-SA	IRT-UA	LogReg
16	-0.21	.005	0.18	0.31	0.32
17	0.18	-.012	0.11	0.32	0.21
18	0.04	-.008	-0.14	0.15	0.07
19	-0.38	.031	0.38	0.55	0.47
20	0.54	-.030	-0.44	0.51	0.54
21	-0.24	.030	0.39	0.54	0.50
22	-0.50	.046	0.00	0.23	0.35
23	0.88	-.066	-0.45	0.45	0.30
24	0.25	-.019	-0.23	0.23	0.12
25	0.48	-.034	-0.36	0.37	0.39
26	-0.20	.033	0.29	0.40	0.54
27	-0.08	.014	-0.08	0.18	0.24
28	0.05	-.006	-0.14	0.26	0.23
29	-0.09	.003	-0.18	0.22	0.78
30	-0.47	.040	0.16	0.22	0.17

Table A.5

DIF Indices for the 30 Items of the GRE-Q with Population Groups Defined by Mathematics Background

Item	MH-D	SIBTEST-b	IRT-SA	IRT-UA	LogReg
1	-0.32	.006	0.11	0.13	0.09
2	-0.06	.006	-0.69	0.78	1.09
3	0.27	-.012	0.60	0.97	0.65
4	-1.43	.145	0.52	0.87	0.76
5	0.13	-.009	-0.11	0.26	0.07
6	-1.42	.122	1.42	1.48	2.93
7	0.38	-.027	-0.15	0.30	0.23
8	0.71	-.050	-0.59	0.59	0.60
9	-0.26	.040	-0.06	0.19	0.40
10	-0.01	.018	-0.17	0.32	0.79
11	0.42	-.003	-0.10	0.50	5.67
12	0.86	-.076	-0.53	0.56	0.57
13	0.42	-.043	-0.40	0.41	0.42
14	0.63	-.061	-0.55	0.59	0.56
15	0.11	-.004	-0.34	0.40	0.12

Table A.5 -- continued.

Item	MH-D	SIBTEST-b	IRT-SA	IRT-UA	LogReg
16	-1.10	.030	0.39	0.45	0.41
17	-0.03	-.015	0.06	0.26	0.11
18	-0.88	.038	1.49	1.79	1.67
19	0.35	-.022	0.06	0.46	0.24
20	-1.49	.140	0.43	0.57	0.75
21	0.46	-.038	0.22	0.75	0.48
22	0.29	-.029	-0.19	0.35	0.13
23	0.26	-.007	-0.30	0.31	0.41
24	-0.36	.025	-0.06	0.10	0.12
25	1.16	-.065	-0.30	0.32	0.29
26	-0.17	.033	-0.03	0.03	0.67
27	0.39	-.024	-0.33	0.35	0.19
28	0.26	-.010	-0.31	0.31	0.11
29	0.58	-.044	-0.26	0.34	0.54
30	-0.29	.039	0.16	0.53	0.78

Table A.6

DIF Indices for the 30 Items of the GRE-Q with Population Groups Defined by Test Anxiety

Item	MH-D	SIBTEST-b	IRT-SA	IRT-UA	LogReg
1	-0.20	.004	-0.08	0.20	0.26
2	0.30	-.009	0.25	0.43	0.14
3	-0.30	.015	0.06	0.27	0.11
4	0.25	-.031	-0.21	0.30	0.12
5	-0.25	.024	0.08	0.24	0.19
6	0.60	-.040	-0.38	0.45	0.42
7	0.25	-.017	0.44	0.52	0.58
8	-0.14	.039	0.07	0.31	0.24
9	-0.18	.013	-0.06	0.22	0.16
10	-0.03	.005	-0.19	0.36	1.37
11	-0.58	.038	0.67	0.74	6.20
12	-0.18	.016	-0.03	0.25	0.09
13	0.07	-.013	-0.09	0.23	0.06
14	-0.40	.025	0.10	0.20	0.14
15	0.42	-.036	-0.24	0.26	0.17

Table A.6 -- continued.

Item	MH-D	SIBTEST-b	IRT-SA	IRT-UA	LogReg
16	-0.82	.022	0.20	0.23	0.20
17	-1.16	.054	0.40	0.41	0.39
18	0.37	-.018	-0.01	0.20	0.12
19	0.08	-.008	0.16	0.27	0.21
20	0.24	-.027	-0.18	0.27	0.28
21	-0.29	.013	-0.26	0.30	0.72
22	0.36	-.030	-0.28	0.33	0.47
23	-0.01	.000	-0.23	0.34	0.05
24	-0.32	.011	-0.03	0.14	0.24
25	-0.36	.006	0.33	0.35	0.24
26	0.24	-.009	-0.01	0.31	0.58
27	0.71	-.066	-0.37	0.41	0.36
28	0.16	-.018	-0.15	0.26	0.19
29	-0.03	-.006	0.19	0.30	0.78
30	-0.07	.003	-0.10	0.25	0.17

Table A.7

Inferential tests of convergent validity coefficients

Methods and Traits	Steiger's z*	p
I. Uniform DIF Procedures		
A.MH and IRT-SA		
Math Bkd. vs. Gender	-0.73	.4654
TA vs. Gender	-1.54	.1236
B.MH and SIBTEST		
Math Bkd. vs. Gender	0.30	.7642
TA vs. Gender	-0.48	.6312
C.IRT-SA and SIBTEST		
Math Bkd. vs. Gender	-0.92	.3576
TA vs. Gender	-1.14	.2542
II. Alternate DIF Procedures		
A.M-H and IRT-UA		
Math Bkd. vs. Gender	-0.30	.7642
TA vs. Gender	-1.05	.2938
B.M-H and SIBTEST		
Math Bkd. vs. Gender	0.04	.9680
TA vs. Gender	-0.07	.9442
C.IRT-SA and SIBTEST		
Math Bkd. vs. Gender	-0.16	.8728
TA vs. Gender	1.65	.0990

Table A.8

Inferential Statistical Tests for the 30 Items of the GRE-Q with Population Groups Defined by Gender

	M-H	SIBTEST	IRT-SA	IRT-UA	LogReg
Item	chi-sq.	z	z	z	chi-sq.
1	10.49**	-3.27**	-1.64	1.63	2.76
2	0.21	0.97	0.30	0.36	0.07
3	2.64	2.28*	2.08*	2.16*	10.93**
4	0.04	-0.56	-0.96	1.95	0.15
5	0.78	-0.74	-1.38	1.38	2.37
6	8.16**	3.18**	1.52	1.08	9.01*
7	0.00	0.27	0.17	0.58	0.28
8	0.25	0.82	-0.68	1.15	7.31*
9	7.86**	3.09**	1.03	1.14	1.98
10	0.44	-0.42	-0.64	1.10	0.04
11	0.07	0.08	-0.03	0.81	4.21
12	0.41	0.77	0.11	1.35	4.76
13	4.85*	-2.31*	-2.85**	4.80**	1.34
14	1.87	-1.67	-1.90	1.90	0.37
15	0.12	-0.69	-1.95	4.27**	0.36

Note. * p < .05
** p < .01

Table A.8 -- continued.

	M-H	SIBTEST	IRT-SA	IRT-UA	LogReg
Item	chi-sq.	z	z	z	chi-sq.
16	0.11	0.22	0.82	1.25	4.97
17	0.13	-0.52	0.46	1.33	2.19
18	0.00	-0.37	-0.52	0.93	0.09
19	0.85	1.29	1.34	1.55	4.64
20	2.47	-1.06	-2.15*	1.33	7.54*
21	0.31	1.24	1.02	1.14	3.12
22	2.26	1.62	0.00	1.00	3.11
23	6.68*	-2.45*	-2.23*	2.26*	2.68
24	0.40	-0.70	-1.97*	1.97*	0.46
25	0.96	-1.61	-1.55	1.48	3.85
26	0.31	1.23	0.69	0.60	2.20
27	0.04	0.46	-0.52	0.54	2.10
28	0.01	-0.21	-0.89	0.99	2.59
29	0.03	0.12	-0.47	0.42	0.01
30	1.77	1.50	0.64	0.69	2.65

Note. * $p < .05$
** $p < .01$

Table A.9

Inferential Statistical Tests for the 30 Items of the GRE-Q with Population Groups Defined by Mathematics Background

	M-H	SIBTEST	IRT-SA	IRT-UA	LogReg
Item	chi-sq.	z	z	z	chi-sq.
1	0.22	0.34	0.18	0.21	0.07
2	0.00	0.27	-0.82	0.86	2.77
3	0.33	-0.53	1.06	1.59	3.50
4	21.29**	4.92**	2.26*	1.94	9.40**
5	0.11	-0.34	-1.38	4.86**	0.03
6	14.95**	4.75**	1.44	1.09	7.68*
7	1.02	-1.07	-0.41	2.70**	0.16
8	4.85*	-1.71	-2.86**	2.66**	5.12*
9	0.55	1.35	-0.36	0.79	6.30*
10	0.00	0.63	-0.62	0.80	3.74
11	0.77	-0.15	-0.07	0.36	8.61*
12	6.70**	-2.61**	-3.79**	2.60**	14.99**
13	1.50	-1.52	-2.98**	2.05*	9.56**
14	2.97	-2.37*	-3.92**	2.58**	18.43**
15	0.07	-0.15	-3.08**	9.13**	1.36

Note. * $p < .05$
** $p < .01$

Table A.9 -- continued.

Item	M-H chi-sq.	SIBTEST z	IRT-SA z	IRT-UA z	LogReg chi-sq.
16	4.95*	1.47	1.43	1.48	2.19
17	0.00	-0.64	0.27	1.14	0.40
18	4.61*	1.67	2.47*	2.72**	25.31**
19	0.73	-0.95	0.23	1.73	2.02
20	20.26**	5.08**	1.52	1.98*	4.40
21	1.30	-1.55	0.58	1.81	4.54
22	0.67	-1.07	0.99	0.96	0.43
23	0.51	-0.24	-1.20	1.65	6.13*
24	0.93	0.96	0.08	0.32	0.24
25	6.17*	-3.16**	-1.14	4.93**	0.26
26	0.18	1.17	-0.06	0.06	3.36
27	1.33	-0.84	-2.13*	0.76	0.43
28	0.56	-0.35	-1.94	2.64**	0.41
29	2.03	-1.58	-0.61	1.32	8.46*
30	0.61	1.35	0.48	1.39	13.79**

Note. * p < .05
** p < .01

Table A.10

Inferential Statistical Tests for the 30 Items of the GRE-Q with Population Groups Defined by Test Anxiety

Item	M-H chi-sq.	SIBTEST z	IRT-SA z	IRT-UA z	LogReg chi-sq.
1	0.06	0.21	-0.15	0.08	0.55
2	0.38	-0.41	0.32	0.48	0.05
3	0.42	0.61	0.12	0.06	0.10
4	0.49	-1.02	-1.28	0.21	1.86
5	0.46	0.84	0.38	0.08	0.78
6	2.32	-1.45	-0.86	0.38	2.87
7	0.36	-0.64	1.09	0.87	4.03
8	0.15	1.26	0.31	0.08	0.38
9	0.21	0.45	-0.38	0.10	0.88
10	0.00	0.17	-0.67	0.96	9.65**
11	1.37	1.73	0.46	0.68	3.26
12	0.23	0.55	-0.16	0.20	0.11
13	0.02	-0.42	-0.71	0.21	0.13
14	1.07	0.90	0.63	0.12	0.20
15	1.28	-1.27	-2.11*	0.23	2.83

Note. * $p < .05$
** $p < .01$

Table A.10 -- continued.

Item	M-H chi-sq.	SIBTEST z	IRT-SA z	IRT-UA z	LogReg chi-sq.
16	2.43	1.07	1.03	0.21	0.21
17	6.79**	2.40*	1.84	0.42	0.95
18	0.64	-0.75	-0.42	0.14	0.36
19	0.02	-0.32	0.23	0.37	1.38
20	0.39	-0.96	-0.88	0.19	1.90
21	0.50	0.51	-0.85	0.46	9.19
22	1.06	-1.03	-1.47	0.31	5.84
23	0.00	0.00	-1.02	0.41	0.10
24	0.70	0.39	-0.30	0.17	4.88
25	0.45	0.26	1.12	0.39	1.11
26	0.37	-0.34	-0.03	0.25	3.43
27	4.45*	-2.21*	-2.46*	0.37	2.64
28	0.18	-0.60	-0.95	0.15	1.60
29	0.00	-0.24	0.42	0.35	5.93
30	0.01	0.12	-0.44	0.10	0.69

Note. * p < .05
** p < .01

Table A.11

Unidimensional Standardized Estimates for the 30-item Test: Exploratory and Cross-Validation Samples

Item	λ	Item	λ
1	0.45 (0.42)	16	0.66 (0.62)
2	0.15 (0.28)	17	0.61 (0.51)
3	0.36 (0.36)	18	0.49 (0.50)
4	0.44 (0.46)	19	0.52 (0.44)
5	0.46 (0.37)	20	0.45 (0.42)
6	0.24 (0.32)	21	0.40 (0.42)
7	0.35 (0.37)	22	0.49 (0.39)
8	0.34 (0.32)	23	0.51 (0.35)
9	0.48 (0.52)	24	0.65 (0.60)
10	0.23 (0.27)	25	0.58 (0.53)
11	0.16 (0.01)	26	0.31 (0.37)
12	0.50 (0.41)	27	0.49 (0.40)
13	0.49 (0.52)	28	0.45 (0.51)
14	0.57 (0.55)	29	0.39 (0.43)
15	0.64 (0.64)	30	0.48 (0.43)

Note. Estimates from the cross-validation sample are in parentheses.

Table A.12

Unidimensional Standardized Estimates for the Amended 26-Item Test: Exploratory and Cross-Validation Samples

Item	*λ*	Item	*λ*
1	0.52 (0.40)	18	0.48 (0.51)
3	0.36 (0.37)	19	0.51 (0.45)
4	0.46 (0.44)	20	0.47 (0.42)
5	0.44 (0.36)	21	0.38 (0.43)
7	0.33 (0.37)	22	0.49 (0.37)
8	0.33 (0.33)	23	0.50 (0.34)
9	0.48 (0.52)	24	0.67 (0.61)
12	0.50 (0.41)	25	0.56 (0.54)
13	0.49 (0.52)	26	0.32 (0.36)
14	0.56 (0.55)	27	0.48 (0.39)
15	0.64 (0.65)	28	0.44 (0.51)
16	0.65 (0.62)	29	0.42 (0.42)
17	0.58 (0.51)	30	0.49 (0.44)

Note. Estimates from the cross validation sample are in parentheses.

Table A.13

DIF Indices Based Upon the 26 Valid Items of the GRE-Q with Population Groups Defined by Gender

Item	MH-D	SIBTEST-b	IRT-SA	IRT-UA	LogReg
1	1.73	-.052	-0.95	0.95	0.92
2	-0.21	.015	.	.	0.52
3	-0.74	.052	1.39	1.54	1.73
4	0.05	-.008	-0.04	0.20	0.02
5	0.24	-.012	-0.29	0.30	0.42
6	-1.03	.080	.	.	5.80
7	-0.12	.005	0.04	0.13	0.14
8	-0.25	.014	-0.10	0.56	0.59
9	-0.91	.091	0.26	0.30	0.39
10	0.23	-.007	.	.	0.39
11	0.13	-.002	.	.	110.45
12	-0.28	.029	0.10	0.43	0.40
13	0.73	-.065	-0.31	0.31	0.27
14	0.45	-.034	-0.25	0.25	0.16
15	0.14	-.021	-0.13	0.25	0.09

Table A.13 -- continued.

Item	MH-D	SIBTEST-b	IRT-SA	IRT-UA	LogReg
16	-0.35	.005	0.20	0.29	0.34
17	0.09	-.009	0.12	0.28	0.24
18	0.14	.003	-0.17	0.36	0.06
19	-0.34	.033	0.34	0.44	0.40
20	0.51	-.029	-0.45	0.61	0.56
21	-0.28	.023	0.38	0.46	0.50
22	-0.49	.047	0.04	0.36	0.39
23	0.78	-.074	-0.31	0.32	0.32
24	0.26	-.027	-0.19	0.19	0.12
25	0.50	-.030	-0.42	0.43	0.39
26	-0.24	.032	0.48	0.61	0.57
27	-0.12	.014	-0.02	0.29	0.26
28	0.09	.000	-0.04	0.40	0.34
29	-0.13	-.001	0.02	0.03	0.14
30	-0.46	.031	0.30	0.39	0.42

Table A.14

DIF Indices Based Upon the 26 Valid Items of the GRE-Q with Population Groups Defined by Mathematics Background

Item	MH-D	SIBTEST-b	IRT-SA	IRT-UA	LogReg
1	-0.27	.002	-0.24	0.27	0.23
2	-0.08	.017	.	.	2.63
3	0.15	-.017	0.35	0.66	0.55
4	-1.49	.141	0.75	0.84	0.77
5	0.04	-.006	-0.16	0.16	0.07
6	-1.43	.126	.	.	11.75
7	0.37	-.043	-0.33	0.33	0.22
8	0.69	-.073	-0.64	0.75	0.60
9	-0.38	.032	0.10	0.44	0.40
10	-0.01	.032	.	.	2.65
11	0.54	-.011	.	.	228.50
12	0.80	-.076	-0.48	0.65	0.54
13	0.39	-.041	-0.37	0.52	0.41
14	0.61	-.060	-0.61	0.74	0.54
15	0.14	-.018	-0.23	0.23	0.05

Table A.14 -- continued.

Item	MH-D	SIBTEST-b	IRT-SA	IRT-UA	LogReg
16	-1.11	.025	0.33	0.36	0.49
17	-0.08	.002	-0.04	0.11	0.12
18	-0.96	.059	1.47	1.72	1.69
19	0.30	-.049	-0.07	0.28	0.22
20	-1.58	.142	0.41	0.87	0.78
21	0.42	-.025	0.02	0.47	0.34
22	0.24	-.041	-0.22	0.22	0.13
23	0.27	-.025	-0.05	0.45	0.47
24	-0.50	.014	-0.05	0.09	0.14
25	1.16	-.064	-0.52	0.52	0.32
26	-0.28	.027	0.38	0.52	0.81
27	0.38	-.030	-0.30	0.30	0.19
28	0.17	-.006	-0.18	0.27	0.21
29	0.60	-.058	0.19	0.68	0.68
30	-0.24	.031	0.47	0.95	0.77

Table A.15

DIF Indices Based Upon the 26 Valid Items of the GRE-Q with Population Groups Defined by Test Anxiety

Item	MH-D	SIBTEST-b	IRT-SA	IRT-UA	LogReg
1	-0.18	.004	-0.18	0.20	0.31
2	0.38	-.012	.	.	0.48
3	-0.24	.001	-0.04	0.10	0.21
4	0.30	-.028	-0.15	0.21	0.27
5	-0.29	.040	0.12	0.12	0.15
6	0.56	-.021	.	.	0.76
7	0.22	-.023	0.36	0.72	0.42
8	-0.16	.028	0.07	0.18	0.30
9	-0.12	.018	0.01	0.03	0.16
10	-0.05	.026	.	.	4.80
11	-0.60	.037	.	.	242.67
12	-0.13	.017	0.02	0.14	0.09
13	0.15	-.011	-0.07	0.12	0.04
14	-0.40	.035	0.12	0.12	0.15
15	0.45	-.036	-0.19	0.19	0.21

Table A.15 -- continued.

Item	MH-D	SIBTEST-b	IRT-SA	IRT-UA	LogReg
16	-0.68	.026	0.18	0.18	0.19
17	-1.12	.068	0.39	0.40	0.37
18	0.27	-.017	-0.18	0.18	0.19
19	0.13	-.011	0.15	0.33	0.21
20	0.23	-.018	-0.21	0.26	0.33
21	-0.22	.032	-0.31	0.57	0.76
22	0.35	-.029	-0.24	0.28	0.37
23	-0.02	.008	-0.14	0.32	0.03
24	-0.37	.005	-0.01	0.21	0.22
25	-0.25	-.003	0.30	0.35	0.25
26	0.19	.002	0.12	0.40	0.62
27	0.74	-.071	-0.34	0.34	0.34
28	0.16	-.010	-0.10	0.14	0.18
29	0.05	-.003	0.37	0.55	0.80
30	-0.05	.006	-0.03	0.05	0.15

Table A.16

Item Inferential Statistical Tests Based Upon the 26 Valid Items with Population Groups Defined by Gender

	M-H	SIBTEST	IRT-SA	IRT-UA	LogReg
Item	chi-sq.	z	z	z	chi-sq.
1	9.98**	-2.87**	-1.83	1.82	3.55
3	2.92	2.46*	2.04*	2.04*	11.64**
4	0.01	-0.28	-0.21	0.71	0.01
5	0.42	-0.41	-1.36	1.18	3.24
7	0.07	0.21	0.10	0.30	0.24
8	0.55	0.46	-0.44	1.36	5.24
9	8.13**	3.09**	1.70	1.68	1.86
12	0.69	0.99	0.64	1.62	5.14
13	4.72*	-2.25*	-2.41*	2.41*	1.38
14	1.52	-1.26	-1.72	0.60	0.42
15	0.12	-0.78	-1.22	1.32	0.81
16	0.45	0.22	0.84	1.19	5.27
17	0.02	-0.39	0.49	1.11	2.39

Note. * $p < .05$
** $p < .01$

Table A.16 -- continued.

Item	M-H chi-sq.	SIBTEST z	IRT-SA z	IRT-UA z	LogReg chi-sq.
18	0.07	0.13	-0.60	0.60	0.01
19	0.65	1.35	1.21	1.32	3.55
20	2.13	-1.05	-2.12*	1.99*	7.00*
21	0.47	0.92	0.97	0.98	3.04
22	2.16	1.64	0.20	1.43	3.56
23	5.12*	-2.73**	-1.53	1.86	4.02
24	0.45	-0.97	-1.65	1.64	0.75
25	0.96	-1.40	-1.79	1.72	4.30
26	0.42	1.16	1.13	1.14	2.52
27	0.10	0.47	-0.13	1.00	2.46
28	0.04	-0.01	-0.27	1.58	5.60
29	0.09	-0.04	-0.06	0.21	0.08
30	1.71	1.17	1.24	1.27	3.24

Note. * $p < .05$
** $p < .01$

Table A.17

Item Inferential Statistical Tests Based Upon the 26 Valid Items with Population Groups Defined by Mathematics Background

	M-H	SIBTEST	IRT-SA	IRT-UA	LogReg
Item	chi-sq.	z	z	z	chi-sq.
1	0.16	0.14	-0.37	0.10	0.45
3	0.08	-0.77	0.61	0.41	2.62
4	23.51**	4.68**	3.14**	2.53*	8.01*
5	0.00	-0.22	-0.62	0.62	0.10
7	0.98	-1.67	-0.87	0.87	0.21
8	4.70*	-2.45*	-3.13**	2.30*	6.03*
9	1.28	1.06	0.51	1.54	5.81
12	6.07*	-2.61**	-3.34**	2.84**	14.19**
13	1.30	-1.42	-2.73**	2.39*	9.89**
14	2.87	-2.25*	-4.27**	4.13**	18.43**
15	0.12	-0.65	-2.10*	2.15*	0.14
16	5.34*	1.24	1.12	1.15	2.61
17	0.01	0.11	-0.17	0.63	0.33

Note. * p < .05
** p < .01

Table A.17 -- continued.

Item	M-H chi-sq.	SIBTEST z	IRT-SA z	IRT-UA z	LogReg chi-sq.
18	5.04*	2.82**	2.31*	2.49*	24.28**
19	0.53	-2.02*	-0.24	1.17	1.76
20	21.95**	5.05**	1.39	2.31*	4.72
21	1.14	-1.02	0.05	1.25	2.30
22	0.47	-1.51	-1.12	1.13	0.25
23	0.55	-0.91	-0.18	0.52	7.00*
24	1.91	0.54	-0.38	0.53	0.40
25	6.07*	-3.21**	-2.05*	2.05*	0.40
26	0.57	0.94	-0.84	1.25	5.12
27	1.29	-1.02	-1.88	1.77	0.78
28	0.23	-0.20	-1.10	1.06	2.19
29	2.25	-2.30*	-0.45	0.41	13.28**
30	0.42	1.09	1.45	1.72	16.47**

Note. * $p < .05$
** $p < .01$

Table A.18

Item Inferential Statistical Tests Based Upon the 26 Valid Items with Population Groups Defined by Test Anxiety

	M-H	SIBTEST	IRT-SA	IRT-UA	LogReg
Item	chi-sq.	z	z	z	chi-sq.
1	0.04	0.26	-0.33	0.35	0.70
3	0.24	0.06	-0.09	0.22	0.42
4	0.74	-0.92	-0.86	0.95	2.61
5	0.64	1.38	0.54	0.54	0.27
7	0.27	-0.87	0.90	1.45	2.20
8	0.19	0.92	0.32	0.50	0.98
9	0.09	0.61	0.05	0.14	0.90
12	0.11	0.56	0.10	0.54	0.14
13	0.14	-0.36	-0.53	0.67	0.07
14	1.12	1.29	0.72	0.68	0.16
15	1.47	-1.26	-1.71	1.93	5.22
16	1.74	1.23	0.90	0.88	0.13
17	6.61**	2.99*	1.73	1.67	0.56

Note. * $p < .05$.
** $p < .01$.

Table A.18 -- continued.

Item	M-H chi-sq.	SIBTEST z	IRT-SA z	IRT-UA z	LogReg chi-sq.
18	0.31	-0.71	-0.73	0.73	0.35
19	0.06	-0.44	0.60	1.13	1.48
20	0.34	-0.63	-0.97	0.86	2.72
21	0.27	1.22	-1.01	1.52	10.32**
22	1.01	-0.98	-1.26	1.04	3.33
23	0.00	0.28	-0.61	0.95	0.03
24	1.04	0.18	-0.13	1.36	4.53
25	0.19	-0.13	1.05	1.11	1.50
26	0.23	0.08	0.29	0.77	3.99
27	4.72*	-2.38*	-2.26*	2.28*	1.72
28	0.18	-0.30	-0.64	0.73	1.59
29	0.00	-0.12	0.80	1.12	7.61*
30	0.00	0.21	-0.14	0.32	0.80

Note. * $p < .05$
** $p < .01$

APPENDIX B
DIFFERENTIAL ITEM FUNCTIONING QUESTIONNAIRE INCLUDING THE REVISED TEST ANXIETY SCALE

DIFFERENTIAL ITEM FUNCTIONING QUESTIONNAIRE

Please carefully bubble in the last five digits of your parents' telephone number.

Please bubble in your age in the first two columns labeled section.

In the second two columns labeled section, bubble in your college classification according to the following criteria:

00 - non-degree
01 - Freshman
02 - Sophomore
03 - Junior
04 - Senior
05 - Master's student
06 - Doctoral student

Directions (questions 1-6): Answer each of the following questions.

1. Your sex:
   a. Female
   b. Male

2. Your ethnic group:
   a. African-American
   b. Asian-American
   c. Hispanic-American
   d. White-American
   e. Other

3. Your college major could be best classified under which of the following:
   a. humanities, fine arts
   b. social science, psychology, education
   c. business
   d. biological sciences
   e. physical sciences, mathematics, engineering

4. Have you successfully completed a college-level calculus course?
   a. yes
   b. no

5. Have you successfully completed a college-level statistics course?
   a. yes
   b. no

6. Have you previously taken the GRE?
   a. yes
   b. no

Directions: The following items refer to how you feel when taking a test. Use the scale below to rate items 7 through 26 in terms of how you feel when taking tests in GENERAL.

1=almost never 2=sometimes 3=often 4=almost always

7. Thinking about my grade in a course interferes with my work on tests......... 1 2 3 4

8. I seem to defeat myself while taking important tests.......................... 1 2 3 4

9. During tests I find myself thinking about the consequences of failing........ 1 2 3 4

10. I start feeling very uneasy just before getting a test paper back............... 1 2 3 4

11. During tests I feel very tense.......... 1 2 3 4

12. I worry a great deal before taking an important exam.......................... 1 2 3 4

13. During tests I find myself thinking
of things unrelated to the material
being tested............................. 1 2 3 4

14. While taking tests, I find myself
thinking how much brighter the other
people are............................... 1 2 3 4

15. I think about current events during
a test................................... 1 2 3 4

16. I get a headache during an important
test..................................... 1 2 3 4

17. While taking a test, I often think
about how difficult it is................ 1 2 3 4

18. I am anxious about tests................ 1 2 3 4

19. While taking tests I sometimes think
about being somewhere else.............. 1 2 3 4

20. During tests I find I am distracted by
thoughts of upcoming events............. 1 2 3 4

21. My mouth feels dry during a test........ 1 2 3 4

22. I sometimes find myself trembling
before or during tests.................. 1 2 3 4

23. While taking a test my muscles are very
tight................................... 1 2 3 4

24. I have difficulty breathing while
taking a test........................... 1 2 3 4

25. During the test I think about how I
should have prepared for the test....... 1 2 3 4

26. I worry before the test because I do
not know what to expect................. 1 2 3 4

APPENDIX C
THE CHIPMAN, MARSHALL, AND SCOTT (1991) INSTRUMENT FOR ESTIMATING MATHEMATICS BACKGROUND

Mathematics Background

1. How many semesters of high school mathematics did you successfully complete?
2. Did you successfully complete a high school calculus course?
3. How many semester credits of college mathematics have you earned?
4. Have you successfully completed a college calculus course?
5. Have you successfully completed a college course in:
    a. physics?
    b. computer science programming?
    c. Engineering?

Composite mathematics background scores were determined by giving students one point for each semester of high school mathematics; two points for successfully completing a high school calculus course; one point for each credit of college mathematics achieved up to a total of ten; two points for successfully completing a college calculus course; and one point each for completing a college course in physics, computer science, or engineering.

REFERENCES

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. Journal of Educational Measurement, 29, 67-91.

Angoff, W.H. (1993). Perspectives on differential item functioning methodology. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 3-23). Hillsdale, NJ: Lawrence Erlbaum.

Angoff, W.H., & Ford, S.F. (1973). Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 10, 95-106.

Baker, F.B. (1981). A criticism of Scheuneman's item bias technique. Journal of Educational Measurement, 18, 59-62.

Benbow, C.P. (1988). Sex differences in mathematical reasoning ability in intellectually talented preadolescents: Their nature, effects, and possible causes. Behavioral and Brain Sciences, 11, 169-232.

Benbow, C.P., & Stanley, J.C. (1980). Sex differences in mathematical ability: Fact or artifact? Science, 210, 1262-1264.

Benson, J., & Bandalos, D. (1992). Second-order confirmatory factor analysis of the Reactions to Tests scale with cross-validation. Multivariate Behavioral Research, 27, 459-487.

Benson, J., & El Zahhar, N. (1994). Further refinement and validation of the Revised Test Anxiety scale. Structural Equation Modeling: A Multidisciplinary Journal, 1, 203-221.

Benson, J., Moulin-Julian, M., Schwarzer, C., Seipp, B., & El Zahhar, N. (1991). Cross-validation of a revised test anxiety scale using multi-national samples. In K. Hagtvet & T.B. Johnson (Eds.), Advances in test anxiety research (Vol. 7, pp. 62-83). Lisse, The Netherlands: Swets & Zeitlinger.

Bentler, P.M., & Bonett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 88, 588-606.

Bishop, Y.M., Fienberg, S.E., & Holland, P.W. (1975). Discrete multivariate analysis: Theory and practice. Cambridge, MA: MIT Press.

Bridgeman, B., & Wendler, C. (1991). Gender differences in predictors of college mathematics performance in college mathematics course grades. Journal of Educational Psychology, 83, 275-284.

Burton, E., & Burton, N.W. (1993). The effects of item screening on test scores and test characteristics. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 321-335). Hillsdale, NJ: Lawrence Erlbaum.

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. Applied Psychological Measurement, 16, 129-147.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Chipman, S.F., Marshall, S.P., & Scott, P.A. (1991). Content effects on word problem performance: A possible source of test bias? American Educational Research Journal, 28, 897-915.

Clauser, B.E., Mazor, K., & Hambleton, R.K. (1991). Influence of criterion variable on the identification of differentially functioning test items using the Mantel-Haenszel statistic. Applied Psychological Measurement, 15, 353-359.

Cohen, A.S., & Kim, S.-H. (1993). A comparison of Lord's chi-square and Raju's area measure in detection of DIF. Applied Psychological Measurement, 17, 39-52.

Cole, N.S., & Moss, P.A. (1989) Bias in test use. In R.L. Linn (Ed.), Educational measurement (3rd Ed. pp. 201-219). New York: American Council in Education/Macmillian.

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart, and Winston.

Darlington, R.B. (1990). Regression and linear models. New York: McGraw-Hill.

Donoghue, J.R., Holland, P.W., & Thayer, D.T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp.137-166). Hillsdale, NJ: Lawrence Erlbaum.

Doolittle, A.E. (1984, April). Interpretation of differenetial item performance accompanied by gender differences in academic background. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Doolittle, A.E. (1985, April). Understanding differential item performance as a consequence of gender differences in academic background. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Doolittle, A.E., & Cleary, T.A. (1987). Gender-based differential item performance in mathematics achievement items. Journal of Educational Measurement, 24, 157-166.

Dorans, N.J., & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum.

Dorans, N.J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. Journal of Educational Measurement, 23, 355-368.

Dorans, N.J., Schmitt, A.P.,& Bleistein, C.A. (1992). The standardization approach to assessing comprehensive differential item functioning. Journal of Educaitonal Measurement, 29, 309-319.

Dweck, C.S. (1986). Motivational processes affecting learning. American Psychologist, 41, 1040-1048.

Eccles, J., Adler, T., & Meece, J.L. (1984). Sex differences in achievement: A test of alternate theories. Journal of Personality and Social Psychology, 46, 26-43.

Educational Testing Service. (1991). Sex, race, ethnicity, and performance on the GRE general test: A technical report. Princeton, NJ: Author.

Educational Testing Service. (1993). GRE: 1993-94 information and registration bulletin. Princeton, NJ: Author.

Elliot, E.S., & Dweck, C.S. (1988). Goals: An approach to motivation and achievement. Journal of Personality and Social Psychology, 54, 5-12.

Elliot, R., & Strenta, A.C. (1988). Effects of improving the reliability of the GPA on prediction generally and on comparative predictions for gender and race particularly. Journal of Educational Measurement, 25, 333-347.

Ethington, C.A., & Wolfle, L.M. (1984). Sex differences in a causal model of mathematics achievement. Journal for Research in Mathematics Education, 15, 361-377.

Everson, H.T., Millsap, R.E., & Rodriguez, C.M. (1991). Isolating gender differences in test anxiety: A confirmatory factor analysis of the Test Anxiety Inventory. Educational and Psychological Measurement, 51, 243-251.

Feingold, A. (1988). Cognitive gender differences are disappearing. American Psychologist, 43, 95-103.

Feingold, A. (1992). Sex differences in variability in intellectual abilities. Review of Educational Research, 62, 61-84.

Fennema, E. (1985). Attribution theory and achievement in mathematics. In S.R. Yussen (Ed.), The growth of reflection in children (pp. 245-265). New York: Academic Press.

Fennema, E., & Petersen, P. (1985). Autonomous learning behavior: A possible explanation of gender-related differences in mathematics. In L.C. Wilkinson & C.B. Marrett (Eds.), Gender influences in classroom interaction (pp. 17-35). New York: Academic Press.

Fennema, E., & Sherman, J. (1977). Sex-related differences in mathematics achievement, spatial visualization and affective factors. American Educational Research Journal, 14, 51-71.

Freedle, R., & Kostin, I. (1990). Item difficulty of four verbal item types and an index of differential item functioning for black and white examinees. Journal of Educational Measurement, 27, 329-343.

Friedman, L. (1989). Mathematics and the gender gap: A meta-analysis of recent studies on sex differences in mathematical tasks. Review of Educational Research, 59, 185-213.

Hackett, G., & Betz, N.E. (1989). An exploration of the mathematics self-efficacy/mathematics performance correspondence. Journal for Research in Mathematics Education, 20, 79-83.

Halpern, D.F. (1992). Sex differences in cognitive abilities (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Hambleton, R.K., & Rogers, H.J. (1989). Detecting potentially biased test items: Comparison of IRT area and Mantel-Haenszel methods. Applied Measurement in Education, 2, 313-334.

Hambleton, R.K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer-Nijhoff.

Harnisch, D.L., & Linn, R.L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. Journal of Educational Measurement, 18, 133-146.

Harris, A.M., & Carlton, S.T. (1993). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. Applied Measurement in Education, 6, 137-151.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. Review of Educational Research, 58, 47-77.

Hill, K., & Wigfield, A. (1984). Test anxiety: A major educational problem and what can be done about it. Elementary School Journal, 85, 105-126.

Holland, P.W., & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.), Test validity (pp.129-145). Hillsdale, NJ: Lawrence Erlbaum.

Holland, P.W., & Wainer, H. (Eds.). (1993). Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum.

Hoover, H.D., & Kolen, M.J. (1984). The reliability of six item bias indices. Applied Psychological Measurement, 8, 173-181.

Hyde, J.S. (1981). How large are cognitive gender differences? American Psychologist, 36, 892-901.

Hyde, J.S., Fennema, E., and Lamon, S.J. (1990). Gender differences in mathematics performance: A meta-analysis. Psychological Bulletin, 107, 139-155.

Joreskog, K.G., Sorbom, D. (1989a). Prelis: A preliminary guide for analysing linear structural relationships. [Computer program]. Chicago: Scientific Software.

Joreskog, K.G., & Sorbom, D. (1989b). LISREL 7: Analysis of linear structural relationships by the method of maximum likelihood [Computer program]. Chicago: Scientific Software.

Kim, S.-H., Cohen, A.S., & Kim, H.-O. (1994, April). An investigation of Lord's procedure for detection of differential item functioning. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Kimball, M.M. (1989). A new perspective on women's math achievement. Psychological Bulletin, 105, 198-214.

Kok, F. (1988). Item bias and test multidimensionality. In R. Langeheine & J. Rost (Eds.), Latent trait and latent class models (pp. 263-275). New York: Plenum Press.

Li, H., & Stout, W. (1993, April). A new procedure for detection of crossing DIF/bias. Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Li, H., & Stout, W. (1994, April). Detecting crossing item bias/DIF: Comparison of logistic regression and crossing SIBTEST procedures. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Licht, B.G., & Dweck, C.S. (1983). Sex differences in achievement orientations: Consequences for academic choices and attainments. In M. Marland (Ed.), Sex differentiation and schooling (pp. 72-97). London: Heinemann Educational Books.

Liebert, R.M., & Morris, L.W. (1967). Cognitive and emotional components of test anxiety: A distinction and some initial data. Psychological Reports, 20, 975-978.

Linn, M.C., DeBenedictis, T., Delucchi, K., Harris, A., & Stage, E. (1987). Gender differences in national assessment of educational progress science items: What does "I don't know" really mean? Journal of Research in Science Teaching, 24, 267-278.

Linn, M.C., & Hyde, J.S. (1989). Gender, mathematics, and science. Educational Researcher, 18 (8), 17-27.

Linn, R.L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp.349-364). Hillsdale, NJ: Lawrence Erlbaum.

Linn, R.L., & Harnisch, D.L. (1981). Interactions between item content and group membership on achievement test items. Journal of Educational Measurement, 18, 109-118.

Linn, R.L., Levine, M.V., Hastings, C.N., & Wardrop, J.L. (1981). Item bias in a test of reading comprehension. Applied Psychological Measurement, 5, 159-173.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. Psychological Reports, 3, 635-694.

Lord, F.M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Maccoby, E.E., & Jacklin, C.N. (1974). The psychology of sex differences. Stanford, CA.: Stanford University.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Marsh, H.W. (1988). Multitrait-multimethod analysis. In J.P. Keeves (Ed.), Educational research, methodology, and measurement: An international handbook (pp. 570-580). New York: Pergamon.

Masters, G.N. (1988). Item discrimination: When more is worse. Journal of Educational Measurment, 25, 15-29.

McCornack, R.L., & McLeod, M.M. (1988). Gender bias in the prediction of college course performance. Journal of Educational Measurement, 25, 321-331.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan.

Miller, M.D., & Linn, R.L. (1988). Invariance of item characteristic functions with variation in instructional coverage. Journal of Educational Measurement, 25, 205-219.

Millsap, R.E., & Everson, H.T. (1993). Methodology review: Statistical approaches for assessing measurement bias. Applied Psychological Measurement, 17, 297-334.

Mislevy, R.J., & Bock, R.D. (1990). BILOG III: Item analysis and test scoring with binary logistic models [Computer program]. Mooresville, IN: Scientific Software.

Moss, P.A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. Reveiw of Educational Research, 62, 229-258.

Muthen, B.O., Kao, C-F, & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. Journal of Educational Measurement, 28, 1-22.

National Center for Education Statistics. (1993). Digest of education statistics 1993. Washington: U.S. Department of Education.

O'Neill, K.A., & McPeek, W.M. (1993). Item and test characteristics that are associated with differential item functioning. In P.W. Holland & H.Wainer (Eds.), Differential item functioning (pp. 255-276). Hillsdale, NJ: Lawrence Erlbaum.

Oshima, T. (1989/1990). The effect of multidimensionality on item bias detection based on item response theory (Doctoral Dissertation, University of Florida, 1989). Dissertation Abstracts International, 51, 829A.

Pajares, F., & Miller, M.D. (1994). Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. Journal of Educational Psychology, 86, 193-203.

Pallas, A.M., & Alexander, K.L. (1983). Sex differences in quantitative SAT performance: New evidence on the differential coursework hypothesis. American Educational Research Journal, 20, 165-182.

Park, D.G., & Lautenschlager, G.J. (1990). Improving IRT item bias detection with iterative linking and ability scale purification. Applied Pscyhological Measurement, 14, 163-173.

Raju, N.S. (1988). The area between two item characteristic curves. Psychometrika, 53, 495-502.

Raju, N.S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. Applied Psychological Measurement, 14, 197-207.

Raju, N.S., Drasgow, F., & Slinde, J.A. (1993). An emprirical comparison of the area methods, Lord's chi-square test, and the Mantel-Haenszel technique for assessing differential item functioning. Educational and Psychological Measurement, 53, 301-314.

Reckase, M.D. (1985, April). The difficulty of test items that measure more than one ability. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Reckase, M.D. (1986, April). The discriminating power of items that measure more than one dimension. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Richardson, F.C., & Woolfolk, R.L. (1980). Mathematics anxiety. In I.G. Sarason (Ed.), Test anxiety: Theory, research, and applications (pp. 271-288). Hillsdale, NJ: Lawrence Erlbaum.

Rosser, P. (1989). Sex bias in college admission tests: Why women lose out. Cambridge, MA: FairTest.

Rudner, L.M. (1977). An approach to biased item identification using latent trait measurement theory. Paper presented at the annual meeting of the American Educational Research Association, New York.

Ryckman, D.B., & Peckham, P. (1987). Gender differences in attributions for success and failure situations across subject areas. Journal of Educational Research, 81, 120-125.

Sarason, I.G. (1980). Introduction to the study of test anxiety. In I.G. Sarason (Ed.), Test anxiety: Theory, research, and applications (pp. 3-14). Hillsdale, NJ: Lawrence Erlbaum.

Sarason, I.G. (1984). Stress, anxiety, and cognitive interference: Reactions to tests. Journal of Personality and Social Psychology, 46, 929-938.

SAS Institute, Inc. (1988). SAS user's guide: Statistics (Version 6.03) [A computer program]. Cary, NC: Author.

Scheuneman, J. (1979). A method of assessing bias in test items. Journal of Educational Measurement, 16, 143-152.

Scheuneman, J.D. (1987). An experimental, exploratory study of causes of bias in test items. Journal of Educational Measurement, 24, 97-118.

Scheuneman, J.D., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. Journal of Educational Measurement, 27, 109-131.

Schmitt, A.P. (1988). Language and cultural characteristics that explain differential item functioning for Hispanic examinees on the Scholastic Aptitude Test. Journal of Educational Measurement, 25, 1-13.

Schmitt, A.P., & Dorans, N.J. (1990). Differential item functioning for minority examinees on the SAT. Journal of Educational Measurement, 27, 67-81.

Schmitt, A.P., Holland, P.W., & Dorans, N.J. (1993). Evaluating hypotheses about differential item functioning. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 281-315). Hillsdale, NJ: Lawrence Erlbaum.

Sells, L. (1978). Mathematics--A critical filter. The Science Teacher, 45, 28-29.

Shealy, R., & Stout, W. (1993a). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. Psychometrika, 58, 159-194.

Shealy, R.T., & Stout, W.F. (1993b). An item response theory model for test bias and differential item functioning. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 197-239). Hillsdale, NJ: Lawrence Erlbaum.

Shepard, L.A., Camilli, G., & Williams, D.M. (1984). Accounting for statistical artifacts in item bias research. Journal of Educational Statistics, 9, 93-128.

Shepard, L.A., Camilli, G., & Williams, D.M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22, 77-105.

Sherman, J. (1983). Factors predicting girls' and boys' enrollment in college preparatory mathematics. Psychology of Women Quarterly, 7, 272-281.

Skaggs, G., & Lissitz, R.W. (1992). The consistency of detecting item bias across different test administrations: Implications of another failure. Journal of Educational Measurement, 29, 227-242.

Spielberger, C.D., Gonzalez, H.P., Taylor, C.J., Algaze, B., & Anton, W.D. (1978). Examination stress and test anxiety. In C.D. Spielberger & I.G.Sarason (Eds.), Stress and anxiety (Vol. 5, pp. 167-191). New York: Hemisphere/Wiley.

Steiger, J.H. (1980). Tests for comparing elements of a correlation matrix. Psychological Bulletin, 87, 245-251.

Stout, W., & Roussos, L. (1992). SIBTEST user manual [Computer program]. Champaign: University of Illinois.

Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. Journal of Educational Measurement, 27, 361-370.

Tatsuoka, K.K., Linn, R.L., Tatsuoka, M.M., & Yamamoto, K. (1988). Differential item functioning resulting from the use of different solution strategies. Journal of Educational Measurement, 25, 301-319.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H.I. Braun (Eds.), Test validity (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum.

Tryon, G.S. (1980). The measurement and treatment of test anxiety. Review Educational Research, 50, 343-372.

Tucker, L.R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. Psychometrika, 50, 253-264.

Uttaro, T., & Millsap, R.E. (1994). Factors influencing the Mantel-Haenszel procedure in the detection of differential item functioning. Applied Psychological Measurement, 18, 15-25.

Wainer, H., & Steinberg, L. S. (1992). Sex differences in performance on the mathematics section of the Scholastic Aptitude Test: A bidirectional validity study. Harvard Educational Review, 62, 323-336.

Wigfield, A., & Eccles, J.S. (1989). Test anxiety in elementary and secondary school students. Educational Psychologist, 24, 159-183.

Wine, J.D. (1980). Cognitive-attentional theory of test anxiety. In I.G. Sarason (Ed.), Test anxiety: Theory, research, and applications (pp. 349-385). Hillsdale, NJ: Lawrence Erlbaum.

Young, J.W. (1991). Gender bias in predicting college academic performance: A new approach using item response theory. Journal of Educational Measurement, 28, 37-47.

Young, J.W. (1994, April). Differential prediction of college grades by gender and ethnicity: A replication study. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. Journal of Educational Measurement, 26, 55-66.

BIOGRAPHICAL SKETCH

Thomas Edward Langenfeld was born in Des Moines, Iowa. He received a Bachelor of Arts in history and education from Iowa State University. Subsequently, he worked as a high school social studies teacher in Storm Lake, Iowa. He later received a Master of Arts in history from the University of Iowa. While teaching at Storm Lake High School, he was cited by the White House Commission on Presidential Scholars for excellence in teaching. After 14 years of public school teaching, he returned to graduate studies to pursue a doctorate. In 1991 he entered the graduate program in research and evaluation methodology in the Department of Foundations of Education at the University of Florida. He received his Ph.D. in 1995 and accepted a position as assistant professor at West Georgia College.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Linda Crocker

Linda Crocker, Chair
Professor of Foundations of
Education

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

James Algina

James Algina
Professor of Foundations of
Education

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Rodman Webb

Rodman Webb
Professor of Foundations of
Education

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Jin Wen Hsu

Jin-Wen Hsu
Assistant Professor of
Foundations of Education

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Marc Mahlios

Marc Mahlios
Professor of Instruction
 and Curriculum

This dissertation was submitted to the Graduate Faculty of the College of Education and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

August, 1995

James Algina

Chairman, Foundations of
 Education

Roderick J McDavis

Dean, College of Education

Dean, Graduate School